DYNAMIC SCHEDULING FOR PRODUCTION SYSTEMS OPERATING IN A RANDOM ENVIRONMENT

THÈSE N^{O} 2825 (2003)

PRÉSENTÉE À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

Institut de production microtechnique

SECTION DE MICROTECHNIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Fabrice DUSONCHET

Diplôme de Mathématiques et informatique, Université de Genève
de nationalité suisse et originaire de Genève (GE)

acceptée sur proposition du jury:

Prof. M.-O. Hongler, directeur de thèse
Prof. R.C. Dalang, rapporteur
Prof. S. Gershwin, rapporteur
Prof. A. Haurie, rapporteur
Prof. J. Jacot, rapporteur

Lausanne, EPFL
2003

Dynamic scheduling for production systems operating in a random environment

Lausanne, EPFL
2003
In part supported by the "Fonds National Suisse pour la Recherche Scientifique"

[...] “Éternellement la science des maîtres passera dans le coeur des disciples, dans un grand silence attentif, comme cette huile rousse de mes collines qui coule du pressoir dans la jarre par un long fil d’or immobile, sans faire de bulles, sans faire de bruit.”
Pagnol (Marcel), Cinématurgie de Paris (Pastorelly).

A Max

1

Abstract in French

La tendance croissante de fabriquer des produits de plus en plus proches du désir de tout un chacun, amène les entreprises d'aujourd'hui à s'équiper de **chaînes de production flexible** (CPF). La flexibilité dans une CPF représente la possibilité qu'a une unique machine de fabriquer différents produits finis, disons N. De plus, cette machine ne peut généralement fabriquer que $M < N$ produits simultanément (nous dirons qu'elle a une "capacité limitée"). Chaque sorte de produit fini répond à une demande spécifique qui possède généralement des fluctuations aléatoires autour d'une valeur moyenne. La présence de ces fluctuations ne permet pas de connaître précisément les demandes futures et impose donc une grande réactivité de l'entreprise afin d'assurer la satisfaction de sa clientèle. En effet, chaque client est une entité égoïste qui a des désirs qu'elle veut, en principe, réaliser sans attendre. Pour faire face à cette difficulté, deux solutions se présentent :

- **La construction de stocks de couverture :**
  La fluctuation *i*) *du flux de production* (due aux enrayages des machines) et *ii*) *du flux de la demande* peux être partiellement absorbée par la construction de stocks de produits finis. Notons que de telles zones de stockage impliquent des coûts non-négligeables surtout en présence d'une demande ayant des désirs très personnalisés. Clairement, le conflit existant entre la volonté de répondre au plus vite à une demande spécifique et celle de minimiser les coûts de production rend l'optimisation du problème compliquée. Généralement, une solution optimale sera un compromis entre ces deux volontés qui se traduira par la construction de stocks de couverture pour certains produits (production sur stock) et non pour d'autres (production à la demande).

- **L'utilisation d'ordonnancement dynamique de production :**
  La présence intrinsèque de fluctuations à l'intérieur d'une chaîne de production implique que des politiques simples d'ordonnancement (comme par exemple des ordres de production déterminés à l'avance qui ne réagissent pas aux aléas possibles du marché) ont peu de chance d'être optimales. Clairement, l'expérience acquise par la connaissance des événements passés et l'observation de l'état présent de la CPF (i.e. le niveau de remplissage des stocks de couverture, la vitesse instantanée de production et la quantité présente des demandes) doivent être à la base de la construction d'une politique optimale d'ordonnancement. Nous comprenons donc qu'une telle politique devra nécessairement s'adapter aux aléas du marché (i.e. une politique d'ordonnancement en temps réel).

Malgré une utilisation grandissante des CPF dans le monde industriel actuel, le problème général de déterminer sa politique optimale d'ordonnancement reste un des problèmes ouverts dans le domaine de la recherche opérationnelle. Dans le but d'apporter quelques réponses à ce problème d'ordonnancement, nous allons discuter dans cette thèse deux

modèles mathématiques connus sous le nom de **"Problème du Bandit à plusieurs bras"** (traduction de "Multi-Armed Bandit problem" (MABP)) et de **"Problème du Bandit Agité"** (traduction de "Restless Bandit problem" (RBP)) à l'aide desquels nous pourrons décrire et discuter la solution de notre CPF. Décrivons brièvement les propriétés principales du problème du Bandit :

Dans sa version classique, le MABP considère une série de N processus stochastiques (aussi appelés projets dans la suite) qui évoluent en parallèle. A chaque instant t, un opérateur de décision (OD) doit engager exactement un des N projets (cette hypothèse reflète la capacité limitée de notre chaîne de production). Le projet engagé génère un coût et évolue en temps. Les projets qui restent inactifs ne coûtent rien et restent figés dans leur état jusqu'à ce qu'ils soient à nouveau engagés (dynamique figée des projets désengagés). Le problème d'ordonnancement dynamique de la MABP consiste à choisir, à chaque instant, lequel des N projets est à engagé afin de minimiser le coût global sur un horizon de temps donné. Une solution optimale pour les MABP a été apportée par Gittins en 1974. Cette politique d'ordonnancement est basée sur la construction d'un ensemble de N indices de priorités (connu sous le nom d'indices de Gittins). Ces indices indiquent l'urgence d'engager chaque projet. Sur la base de ces indices, l'ordonnancement optimal est comme suit :

"A chaque instant, engager le projet possédant la plus petite valeur d'indice de Gittins"

Pour pouvoir utiliser le formalisme des MABP afin de modéliser notre problème de chaîne de production flexible, nous devons nous débarrasser de la supposition que la dynamique des projets désengagés est figée. En effet, nous devons autoriser la situation suivante :

a) Les projets qui sont désengagés évoluent en temps et génèrent des coûts. Cette généralisation est nécessaire pour notre CPF, en tenant compte que la demande des produits non fabriqués actuellement continue d'arriver. De plus, leurs stocks génèrent des coûts.

L'hypothèse *a)* nous amène à étudier le problème connu sous le nom de "Bandit Agité" (RBP) pour lequel la politique optimale n'est pas encore connue. Comme il est écrit dans la (rare) littérature existante sur les RBP, une généralisation naïve de la politique d'indices de priorité conduit à un résultat loin de l'optimum. Néanmoins, des heuristiques (sous-optimales) apportant un résultat proche de l'optimum peuvent être obtenues à partir d'une **généralisation adéquate des Indices de Priorité**. L'utilisation de tels indices a l'avantage de fournir des heuristiques simples, ce qui est une condition essentielle pour pouvoir les appliquer aux problèmes industriels que nous avons à l'esprit. Un des buts rattachés à notre étude des CPF sera donc de construire de telles heuristiques. En particulier, nous nous efforcerons de trouver des modèles simples dont la politique optimale d'ordonnancement peut être dérivée explicitement. Ceci nous aidera à développer la perception nécessaire à la construction d'heuristiques fiables, applicables aux problèmes généraux.

En conséquence, l'approche adoptée dans cette thèse est la suivante :

i) Construire des classes de problèmes de Bandits (classiques et agités) pour lesquels la politique optimale d'ordonnancement peut être dérivée explicitement.

ii) Développer une heuristique permettant d'approcher la solution optimale pour le problème d'une chaîne de production flexible et tester son efficacité en la comparant aux politiques optimales dérivées au point *ii)*.

De plus, notons que la flexibilité dans les CPF génère la plupart du temps des coûts et du temps de set up lors du changement entre un type de production et un autre (tel que le coût de la main d'oeuvre supplémentaire nécessaire au lavage des machines) nous allons donc aussi :

iii) construire une classe de Bandits classiques avec pénalités de set up pour laquelle la politique d'ordonnancement peut être dérivée explicitement.

iv) Développer une heuristique permettant d'approcher la solution optimale pour le problème général des MABP avec des pénalités de set up et tester son efficacité en la comparant à la politique optimale dérivée au point *iii*).

En résumer nos contributions sont les suivantes :

- **Calcul explicite de l'indice de Gittins pour une nouvelle classe de MABP (sans coût de set up) :**
  Nous avons été capables de construire explicitement la forme de l'indice de Gittins quand l'évolution sous-jacente est donnée par un processus déterministe par morceau. Un tel processus est intrinsèquement non-Markovien. Cette classe de MABP est une des première classes non-Markovienne que nous pouvons trouver dans la littérature pour laquelle l'indice de Gittins peut être calculé explicitement (M.-O. Hongler and F. Dusonchet, "Optimal stopping and Gittins indices for piecewise deterministic evolution process", Discrete Events Systems (2001) (**11**), 235–248).

- **Discussion explicite du problème du Bandit Agité :**
  Nous avons étudié plusieurs dynamiques stochastiques sous-jacentes directement utiles pour le contexte industriel (e.g. processus de diffusion et processus de naissance et de mort). Nous avons obtenu des formules explicites d'indices de priorité généralisés et l'heuristique en découlant a été comparée avec la politique optimale construite numériquement par Ha (Oper. Res. **45**, 1994, 42-53). Finalement, en se basant sur les RBP, nous avons proposé une politique d'ordonnancement efficace pour une machine flexible pouvant fabriquer plusieurs types de produits, lorsque les pénalités de set up peuvent être négligées (F. Dusonchet and M.-O. Hongler, "Continuous Time Restless Bandit and Dynamic Scheduling for Make-to-Stock Production", accepted for publication by IEEE Trans. on Robo. and Auto., (2003)).

- **Construction d'une heuristique d'ordonnancement sous-optimal pour le problème des MABP avec pénalités de set up :**
  Tenir compte des pénalités induites par les set up accroît singulièrement la difficulté de la recherche d'une politique optimale. De ce fait, il existe très peu de résultats sur ces sujets et la recherche de la politique optimale des problèmes de décision avec pénalités de changement reste un problème peu exploré. Pour la fin de cette thèse, nous avons donc concentré nos efforts sur le problème des MABP avec coûts et temps de set up. Un progrès significatif a été obtenu étant donné que nous avons pu construire une nouvelle classe de MABP avec pénalités de set up pour laquelle la politique optimale d'ordonnancement peut être calculée explicitement. Sur la base de cette politique optimale, nous avons pu proposer une heuristique approchant la solution générale des MABP avec pénalités de set up. Cette heuristique est basée sur une généralisation de la politique d'indices de priorité (F. Dusonchet and M.-O. Hongler, "Optimal Policy for Deteriorating Two-Armed Bandit Problems with Switching Costs", accepted by Automatica, (2003)).

2

Abstract in German

Dem zunehmenden Bedürfnis der Industrie Produkte mit immer mehr Rücksicht auf die fluktuierende (d.h. zufällig schwankende) Nachfrage herzustellen, begegnet die industrielle Produktion mit dem Konzept flexibler Fabrikationssysteme (FFS). Dabei versteht man in diesem Zusammenhang unter Flexibilität die Fähigkeit einer einzelnen Produktionszelle mehrere (sagen wir N) verschiedene Teil- oder Fertigprodukte (folgend kurz Produkte genannt) herzustellen. Gewöhnlich ist die Produktionskapazität limitiert, so dass höchstens $K < N$ verschiedene Produkte gleichzeitig gefertigt werden können. Jedes Produktionsgut sieht sich dabei einem entsprechenden Absatzmarkt gegenüber der, wie bereits erwähnt, zufälligen Fluktuationen unterliegt. Um auf diese Schwankungen rasch zu reagieren sind effiziente Produktionsstrategien für FFS unabdingbar. Solche Strategien beinhalten im wesentlichen zwei Punkte:

1. **Endproduktlager.** Fluktuationen im Produktionsfluss (auf Grund fehleranfälliger Produktionsinstallationen) und in der Nachfrage können durch einen entsprechenden Lagerbestand fertiger Produkte (LFP) partiell aufgefangen werden. Eine Strategie die sich auf ein Endproduktlager abstützt, muss das durch den Lagerbestand gebundene Kapital mit den Vorteilen einer dadurch gewonnenen hohen Reaktivität in ein günstiges Verhältnis stellen. Das resultierende Optimierungsproblem ist komplex und die optimale Lösung beinhaltet in der Regel hybride Produktionsregeln (d.h. verschiedene Produkte verlangen verschiedene Produktionsstrategien).

2. **Dynamische Strategien.** Die naturgemäss vorhandenen Fluktuationen haben zur Folge, dass einfache, deterministische Produktionsregeln (wie z.B., "fertige periodisch jedes Produkt während einer gewissen Zeitspanne") einen sehr bescheidenen Leistungsnachweis erbringen. Eine gute Strategie bedient sich zum Einen, vorhandenen Erfahrungswerten und zum Anderen, dem aktuellen Zustand des Produktionssystems und der Nachfrage (z.B., Information über LFP, Produktionsflüsse und Bestelleingänge). Eine optimale Strategie wird demzufolge einen zeitabhängigen adaptiven Charakter aufweisen müssen ("Real-time" Strategie).

Trotz seiner unbestrittenen Relevanz für industrielle Anwendungen bilden real-time Optimierungsprobleme für FFS nach wie vor ein weites, keineswegs abgeschlossenes Forschungsfeld. Die vorliegende Doktorarbeit platziert sich in diesem Arbeitsfeld mit zwei mathematischen Modellen bekannt unter den Namen "Multi-Armed Bandit Problems" (MABP) und "Restless Bandit Problem" (RBP). Dieser "Bandit"-Formalismus erlaubt die Beschreibung einer FFS. Er sei hier folgend inhaltlich kurz zusammengefasst.

In seiner Grundversion besteht das MABP aus N stochastischen Prozessen (auch "Projekte" genannt), die sich zeitlich parallel entwickeln. Zu jedem Zeitpunkt t kann höchstens ein (Teil)-Projekt aktiv sein (sprich, hergestellt werden). Die Herstellung eines Projekts generiert entsprechende Herstellungskosten doch wird angenommen, dass während dieser Herstellzeit die übrigen Projekte weder Kosten verursachen noch ihren Zustand ändern

(nicht aktive Projekte sind in ihrem Ist-Zustand "erstarrt"). Das dynamische Grundproblem besteht nun darin zu jedem Zeitpunkt aufs neue zu entscheiden welches Projekt herzustellen ist, so dass die globalen Produktionskosten während eines gegebenen Produktionszeitraums minimiert werden. Dieses Grundproblem wurde 1974 von Gittins exakt gelöst. Seine Lösung basiert auf der Konstruktion von N Index-Funktionen (Gittins-Index) welche jedem Produkt eine "Dringlichkeitsfunktion" zuweist und mit deren Hilfe die optimale dynamische Produktionsstrategie wie folgt zusammengefasst werden kann:

"Stelle zu jedem Zeitpunkt das Produkt mit kleinster Dringlichkeitsstufe her."

Um den MABP-Formalismus auf industrielle Produktionssysteme anwenden zu können muss die oben genannte Grundhypothese – nicht-aktive Projekte sind "erstarrt" – aufgeweicht werden. Tatsächlich sollte sie wie folgt verallgemeinert werden:

(H) Auch nicht aktive Projekte können ihren Zustand ändern und Kosten verursachen.

Die Verallgemeinerung ist notwendig um der sich verändernden Nachfrage eines (nicht notwendigerweise aktiven) Projekts zu begegnen und seine Lagerkosten zu berücksichtigen. Lagerkosten und allgemein Kosten die nicht direkt produktionsbedingt sind, werden folgend kurz als Zusatzkosten angesprochen.
Diese Annahme veranlasste uns die in diesem Fall adäquateren "Restless-Bandit" Probleme anzugehen für die jedoch im allgemeinem noch keine optimalen Lösungen vorliegen. Die spärliche Literatur zum mathematischen Problem beinhaltet im wesentlichen ein "negatives" Resultat welches zeigt, dass eine naive Erweiterung der oben erwähnten Index-Lösung zwangshalber sub-optimal ist. Trotzdem können durch **generalisierte Index-Strategien** gute Produktionsregeln aufgestellt werden die nahe der optimalen Lösung sind. Der entscheidende, die Sub-Optimalität überwiegende Vorteil von Index-Strategien ist ihre sehr einfache Implementierung was mit Blick auf unsere industrielle Anwendung in der Tat essentiell ist. Die Konstruktion einfacher, optimal lösbarer "Bandit" Probleme soll dabei helfen, die sub-optimalen generalisierten Index-Strategien auf ihre Verwendbarkeit hin zu untersuchen. Entsprechend wurde der vorliegenden Arbeit folgendes Programm zu Grunde gelegt:

(i) Konstruktion einer Klasse von explizit lösbaren Bandit Problemen.

(ii) Entwicklung heuristischer Methoden für FFS und Studium ihrer Verwendbarkeit auf Grund numerischer Simulationen.

Darüber hinaus soll mit Blick auf die industrielle Anwendung im Bereich FFS die Arbeitshypothese (H) wie folgt integriert werden:

(iii) Konstruktion einer Klasse von explizit lösbaren MABP mit Zusatzkosten.

(iv) Entwicklung heuristischer Methoden für MABP mit Zusatzkosten und Studium ihrer Verwendbarkeit basierend auf einem Vergleich mit den explizit konstruierten Modellen in (iii).

Unser Beitrag zu diesem Forschungsprogramm ist:

- **Explizite Berechnung des Gittins-Index für MABP ohne Zusatzkosten.** Für eine nicht-Markovsche, stückweise deterministische Projektdynamik wurden die Gittins-Index Funktionen explizit berechnet. Es ist dies das erste nicht-Markovsche Model mit expliziter Lösung und bereits publiziert unter "Optimal Stopping and Gittins indices for picewise deterministic evolution process", M.-O. Hongler und F. Dusonchet in "Discrete Events Systems" (2001) Vol. 11, 235-248.

- **Explizite Teillösungen für "Restless-MAB" Prozesse.**
  Verschiedene, für die industrielle Produktion relevante, stochastische Projektdynamiken (z.B. Diffusionsprozesse) wurden untersucht. Wir konstruierten explizite (generalisierte) Index-Funktionen und verglichen die resultierende Produktionsstrategie mit den numerischen Resultaten der implizit gegebenen optimalen Lösungen von A. Ha (Oper. Res. 45, 1994, 42-53). Schlussendlich wurde mit Hilfe des "Restless-Bandit" Formalismus eine heuristische Lösung für das "Multi-Item Make-to-Stock" Produktionsproblem vorgeschlagen. Dieser Beitrag ist zur Publikation in IEEE "Transactions on Robotics and Automation" (2003) unter dem Titel "Continuous Time Restless Bandit and Dynamic Scheduling for Make-to-Stock Production" (F. Dusonchet und M.-O. Hongler) freigegeben.

- **Konstruktion einer Heuristik für MABP mit Setup Kosten.**
  Werden zusätzlich Setup Kosten (z.B., die Kosten die beim Aktivieren eines passiven Projekts anfallen) in das Optimierungsproblem aufgenommen, wird ein singuläres Lösungsverhalten immer wahrscheinlicher und eine allfällige Lösung wird sehr kompliziert ausfallen. Es ist deshalb nicht verwunderlich, dass bis heute kaum Resultate existieren die solche Kosten miteinbeziehen. Unsere Anstrengungen konzentrierten sich auf MAB-Probleme mit Setup Kosten und ein signifikanter Fortschritt wurde durch die Identifizierung einer Klasse von MABP erzielt für die die optimale Strategie durch rekursive Konstruktion explizit angegeben werden kann. Mit Hilfe dieser speziellen optimalen Lösung wurde anschliessend eine heuristische Strategie für das generelle MAB-Problem mit Zusatzkosten abgeleitet. Dieser Beitrag ist zur Publikation in "Automatica" (2003) unter dem Titel "Optimal Policy for Deteriorating Two-Armed Bandit Problems with Switching Costs" (F. Dusonchet und M.-O. Hongler) freigegeben.

3

Abstract in English

Due to the steady tendency to propose highly customized products and to respond to volatile (i.e random) demands, **Flexible Manufacturing Systems** (FMS) are now present in most shopfloors. In this thesis, flexibility in a FMS is understood as the ability of a single production cell to deliver several different types of items, say N. The production capacity is usually limited in the sense that only one or $K < N$ type(s) of items can be simultaneously produced. Each type of item faces a specific market demand which often shows **random fluctuations**. To insure a high reactivity when facing such random demands, efficient production rules for the FMS are mandatory. These rules include in particular two generic entities, namely:

- **Finished Goods Inventories**. The fluctuations of *i*) *the production flows*, (due to failure-prone machines) and *ii*) *the demand flows*, can be partly absorbed by the presence of Finished Good Inventories (FIG). Such storage zones incur costs especially when serving highly customized demands. Clearly, the balance between the advantage of high reactivity on the one hand and storage costs on the other introduces complex optimization issues. The optimal solution will generally include hybrid production rules, i.e. certain types of products optimally require FIG (*we call this strategy make-to-stock production*), while other types require no FIG (*we call this strategy make-to-order production*).

- **Dynamic Scheduling rules**. The intrinsic presence of fluctuations implies that simple deterministic scheduling rules (as for example deterministic polling rules which produce each type j of items periodically during a fixed period of time T_j) may lead to a very poor performance. Clearly, an optimal production schedule will be based on both past experience and observation of the present state of the system (i.e the populations of the FIG and the instantaneous rates of the demand and production flows). Hence, any optimal scheduling rule will necessarily present a time adaptive character (i.e. real time scheduling rules).

In spite of a growing usage of FMS in industry, the general problem of determining the **optimal dynamic scheduling** of flexible manufacturing systems remains, in its full generality, an open issue of operations research. In order to give some answers to the question of optimal scheduling, the present thesis will discuss two mathematical models known as "**Multi-Armed Bandit Problem**" (MABP) and "**Restless Bandit Problem**" (RBP) in terms of which the FMS can be modeled. Let us briefly recall the salient features of the Bandit formalization.

In its basic version, the MABP considers a series of N stochastic processes (also called *projects*) evolving in parallel. At each time t, a decision maker (DM) can engage at most one project (this feature reflects the limited resource property). The engaged project

generates an instantaneous cost while the disengaged projects incur no cost and remain fixed ("frozen" dynamic rule of the disengaged projects). The optimal scheduling problem in the MABP consists in choosing at each time t which project to engage in order to minimize the global cost over a given time horizon. An exact solution of the basic MABP has been given in 1974 by Gittins. The solution is based on the construction of a set of N priority indices (the so-called Gittins indices) which assign to each project an "urgency function" in terms of which the optimal dynamic scheduling reads:

"At each instant, engage the item exhibiting the smallest priority index value".

To use the MABP formalization in the production engineering context, the basic hypothesis of the "frozen" dynamics needs to be loosened. It is indeed mandatory to allow for the following features:

a) Disengaged projects evolve in time and do incur costs. This generalization is necessary to reflect the fact that demands for items not currently produced continue to accrue and their FGI obviously also incurs costs.

The assumptions in *a*) lead us to study the so-called "Restless Bandit problems" for which the optimal scheduling rule is yet unknown. As it has been noted in the (scarce) literature available, a naive generalization of the priority indices will definitely not yield the optimal rule. However, close to optimal solutions can still be expressed in terms of suitably **generalized priority indices**. The use of such indices has the determining advantage of leading to very simple –though sub-optimal– scheduling policies. This is indeed an essential feature of the production applications we have in mind. Hence, the construction of simple solvable models of optimal scheduling rules, will help to develop the perception needed to construct reliable heuristics applicable to the general problems.

Accordingly, the general approach adopted in the present thesis is:

i) Construct explicitly solvable classes of Bandit problems (Classical and Restless).

ii) Develop a heuristic to approach the problem of FMS and tests its validity on the basis of simulation studies.

Moreover, as flexibility in a FMS generally generate **setup penalties** (such as the need for additional workforce incurring additional costs or cleaning operations imposing switching time delay) we will further:

iii) Construct an explicitly solvable class of MABP with setup penalties.

iv) Develop a heuristic to approach the general problem of MABP with setup costs and test its validity on the basis of the simple models introduced in iii).

Our original contributions are:

- **Explicit computation of the Gittins index for MABP (without switching penalties).**
  We were able to compute explicitly the form of the Gittins indices when the evolution is given by a piecewise deterministic process which is intrinsically non-Markovian. This is among the few classes of non-Markovian examples in the literature for which the Gittins indices can be computed explicitly (M.-O. Hongler and F. Dusonchet, "Optimal stopping and Gittins indices for piecewise deterministic evolution process", Discrete Events Systems (2001) (**11**), 235–248).

- **Explicit treatment of the Restless Multi-Armed Bandit process.**
  We studied several underlying random dynamics relevant for the production engineering context, (e.g. diffusion processes as well as birth and death processes). We obtained explicit generalized priority indices and the resulting dynamic scheduling was compared with exact results derived numerically by A. Ha, (Oper. Res. **45**, 1994, 42-53). Finally, using the RBP, we propose a sub-optimal heuristic solving the multi-items make-to-stock production problem (F. Dusonchet and M.-O. Hongler, "Continuous Time Restless Bandit and Dynamic Scheduling for Make-to-Stock Production", accepted for publication by IEEE Trans. on Robo. and Auto., (2003)).

- **Construction of a sub-optimal heuristic for the MABP with setup penalties.**
  Adding setup penalties to optimal control problems singularly increases the complexity of the solution. Therefore, very few results presently exist and the optimal decision problem with setups remains mostly unexplored. Our effort concentrates on the MABP with setup penalties and significant progress has been made by constructing a new class of MABP with setups for which the optimal policy can be explicitly constructed by recursion. Using this optimal derivation, we then propose a heuristic, approaching the optimal policy for general MABP with switching penalties (F. Dusonchet and M.-O. Hongler, "Optimal Policy for Deteriorating Two-Armed Bandit Problems with Switching Costs", accepted by Automatica, (2003)).

Preface

This thesis is about scheduling optimization in an industrial context. Like any subject, scheduling optimization lends itself to different treatments, ranging from a very pragmatic and problem centered approach to a more generalized quest for knowledge. And the very existence of these two polarities has been one of the key hurdles I had to overcome. All through the years I spent writing this thesis, I recurrently faced the dilemma:

"Should I try to come up with practical solutions to specific industrial scheduling problems or should I transcend specifics in order to gain deeper and more generic understanding of the generalized scheduling problem?"

At one point I actually thought I could avoid a definite choice and strike instead a fine balance between the two contradictory approaches. But I must admit however that it would be perhaps pretentious and certainly misleading if I claimed to have achieved such a goal. In the end the theoretical slant prevailed over the practical, not least because I felt that gaining broad knowledge was more in keeping with the very nature of a PhD thesis. For it can safely be assumed that if we are able to fully understand and model a process, in due course such insight will lead to innovative practical solutions and serve as a platform for knowledge sharing (through articles, seminars, etc) with the industrial and the scientific community at large.

Given my choice, I would not be surprised if many practitioners of the scheduling art may find the heuristic approach both frustrating and unsatisfactory. That would be par for the course. I am conscious that running a production line is not like sipping tea at a mid-summer garden party. On the contrary, it is a taxing challenge to keep production rolling, day in day out, month after month and year after year, constantly facing a variety of problems - most of them urgent, some of them vital - with no slack time whatever for dreaming up lofty theories and innovative solutions. Product line management is all about the here and now. To paraphrase a famous Ross Perot [1] saying:

"*...on the shop floor, if you see a rattlesnake you grab a shovel and kill it. You don't walk off to invent the new snake vaccine.*"

Men who keep the supply chain constantly stocked as per manufacturing requirements have all my admiration. And if some of them are reluctant to tackle the mathematics and modeling that would give them a better understanding of the dragon they are trying to slay, this is perfectly understandable. While it is indeed difficult to justify that a production manager should perform the detached work of a scientist, this limitation does not in any way imply that the theoretical view should always be looked upon with suspicion. Indeed we can safely assume that well managed R&D (Research and Development) is often the key to competing successfully in a crowded marketplace. And

[1] founder of EDS and one-time presidential US candidate

when it comes to R&D, both scientists and industrial decision makers agree on at least one fundamental point, i.e. that "**...building a sound model is the starting point for understanding and, ultimately improving complex processes.**" Once a model is found, solutions can then be delivered in the shape of

a) analytical study,

b) numerical computation or

c) simulation tools.

Whatever choices one may want to emphasize from the above list, it should be crystal clear that time and effort should be spent up-front on analyzing and modeling the specifics of the given problem. Unfortunately, human nature being what it is, it is not surprising that a number of ads in the specialized press offer quick and inexpensive fixes of "Simulations without Modeling" [8]. One can but assume that such unscrupulous promotions are initiated by fast buck artists, fishing for naive prospects at the high-tech end of the engineering marketplace.

Needless to say, it has not been my aim to cut any such corners in this thesis. Even though our models may at times oversimplify reality (but in a way isn't that what modeling is all about?), we have certainly made every possible effort to insure that they contain all the essential ingredients that make up the underlying reality of scheduling problems. Furthermore, whatever simplifications we may have introduced and whatever assumptions we may have built into our models, have been clearly enunciated and described. Thus our whole treatment of the subject matter has been totally transparent and is open to criticism. Finally, beyond all the math and the modeling, we sincerely hope that this thesis will be first and foremost thought provoking.

At its deepest level, reality is mathematical in nature. Pythagore.

Contents

Every section which present a new sesult
has its title beginning with a star $(\circledast)$

1

Introduction

1.1 Optimal Decision Problem

Which is the best course of action to take given what is already known about a situation? This natural question is in the center of "Decision Theory". Decision theory is an important subject of the mathematics, taking into account three categories of problems, namely:

a) **Decision-making in a deterministic environment.**
In this case, each action taken has a corresponding single outcome which is exactly known in advance.

b) **Decision-making in a random environment (i.e. risky or uncertain).**
Here, several reactions may follow a single action taken. Each particular reaction will depend directly on the state of the environment and its consequences will follow a law of probability which may or may not be known *a priori*. (Take for example the result of throwing a dice, which will be a value between 1 and 6 with a probability of $\frac{1}{6}$).

c) **Decision-making in a situation of conflict.**
In this case, it is the behaviour of the adversary which creates the incertitude and not the environment itself. This class of decision problem is known under the name of "Games theory".

In this thesis, we decided to focus our attention on the problems belonging to category *b*). Specifically, we will study a mathematical formalism which allows a choice between actions yielding immediate reward and others (such as acquiring information or skill, or preparing the ground) whose benefit will appear only later but potentially may lead to a higher global reward. A common characteristic of these problems is the random nature of the dynamics of the underlying environment. This random behaviour impose that the optimal actions cannot be determined simply as they would be for deterministic evolutions. In general the environment behaviour drastically influences the decision-making. Problems of this type belong to the class of "sequential decision problems". Here "sequential" suggests that a Decision Maker (DM) must make his decision by relying only on his past experience and that no premonition is allowed. Moreover he may change his mind at any time in a sequential manner. Dealing with sequential decision-making problem, we can distinguish between two main issues:

a) System identification: How to use the information obtained from the observed outputs to reduce the uncertainty about the system's behaviour in the future.

b) Stochastic control: Select feedback control actions, (i.e. closed loop reaction), to optimize the values of a relevant cost function.

In order to illustrate both cases *a*) and *b*) above let us consider the industrial world: Most manufacturing firms are large, complex systems characterized by several decision subsystems, such as finance, personnel, marketing and operations. They commonly have a number of plants and warehouses and produce a large number of different products using a wide variety of machines and equipment. Moreover, these systems are subject to discrete events such as the construction of new facilities, the purchase of new equipment and scrappage of old, machine setups, failures and repairs and the introduction of new products. These events could be deterministic or stochastic. To be efficient, management must be aware of and react to these events. We thus understand the importance of constructing an efficient model of the manufacture device and efficient scheduling heuristics in order to achieve a very rapid (ideally instantaneous) response to random demands while maintaining production costs as low as possible.

Because of the large size of these firms and the occurrence of such events, obtaining exact models and optimal policies to run these systems is nearly impossible both theoretically and computationally. Let us look, as an example, at firms producing watches. In such manufacture, the tendency is to propose very customized products matching each individual desire. To match this requirement, the engineer usually constructs flexible production lines, each of which is able to produce many different types of finished goods. Usually only a single type of product can be delivered at a given instant and therefore specifications of production plans (i.e. a proper scheduling for the type of production to engage) are mandatory. At each decision time, the production schedule will be based on both available past experience and present information regarding inventories, production and demand flows. Practice shows that actual flexible production and/or assembly lines always involve setup penalties such as cleaning operations. During these setups, the production is stopped and an additional workforce is usually needed. Hence, these setups are costly and it is therefore very important to schedule the precise periods for changing the production type.

Although flexible production is well implanted in actual manufacturing systems, we can observe that the optimal scheduling of their production flow is generally not known. In industry, the Decision Maker (DM) schedules his installation using very simple decision rules. In most cases, these simple rules perform well enough for the survival of the firm, but are far from the optimal strategies. Therefore, every step towards the understanding of optimal scheduling is welcome and will help to develop more efficient decision rules. Accordingly, this thesis will study the dynamic scheduling of Flexible Manufacturing Systems (FMS) operating in a random environment.

One of the most important methods in dealing with the optimization of large, complex systems such as FMS is that of hierarchical decomposition. The idea is to reduce the overall complex problem into manageable approximate ones, to solve them and to construct a solution of the original problem from the solution of the simpler ones. This idea has been identified as particularly fruitful when applied in the area of Operations Research (see [44]) but is also a general idea used to discuss any difficult problem. Indeed, we can cite, for example, Descartes who, in his "Discourse on Method", already introduced this concept as follows:

[...] *"Instead of the great number of precepts of which logic is composed, I believed that the four following would prove perfectly sufficient for me, provided I took the firm and unwavering resolution never in a single instance to fail in observing them.*

The first was never to accept anything for true which I did not clearly know to be such; that is to say, carefully to avoid precipitancy and prejudice, and to comprise nothing more in my judgement than what was presented to my mind so clearly and distinctly as to exclude all ground of doubt.

The second, to divide each of the difficulties under examination into as many parts as possible, and as might be necessary for its adequate solution.

The third, to conduct my thoughts in such order that, by commencing with objects the simplest and easiest to know, I might ascend by little and little, and, as it were, step by step, to the knowledge of the more complex; assigning in thought a certain order even to those objects which in their own nature do not stand in a relation of antecedence and sequence.

And the last, in every case to make enumerations so complete, and reviews so general, that I might be assured that nothing was omitted."

Descartes "The Discourse on the Method" [10].

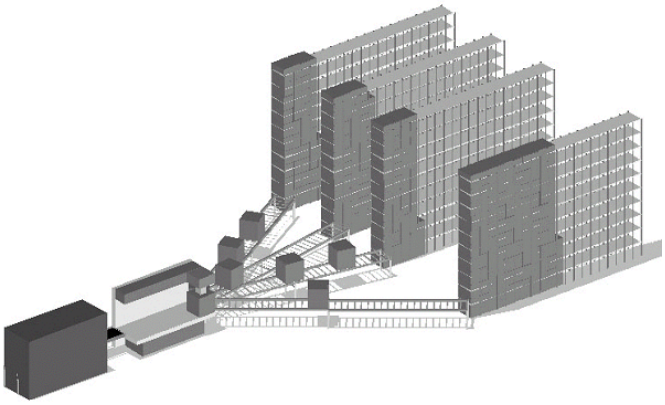

Fig. 1.1. "Make-to-stock" production facility (MSP).

Let us then make some approximations in our initial FMS in order to obtain a tractable problem. From the point of view of external demand, we may consider our FMS as a single flexible production entity with a set of finished goods inventories (i.e. a "make-to-stock" production (MSP) facility with multiple products, see figure 1.1). Specifically, the MSP consists of a single production unit which is able to manufacture N types of different products, however only one at a time. Each finished product is placed in its respective inventory which services an exogenous demand. When demand for a product arises, it is either satisfied immediately from on-hand inventory, if available, or is backordered otherwise (i.e. the demand waits until the desired good is produced). In a MSP we normally assume that there is plenty of raw material at the entrance of the production unit so that the machine can always be supplied with the parts needed for its work. Moreover we normally assume that the production time for each good is fixed but that the demand for this good has fluctuations around its mean value. These assumptions are not always proven accurate in the real flexible production line. Indeed, the supplier may have delays causing the production of one type of goods to stop because of raw material shortage. Moreover, the production line being generally composed of several workstations, the time needed for the production of one good may be random. Therefore considering a make-

to-stock production instead of the actual flexible production line is a first approximation.

Although the MSP is a simplified problem, both difficulties *a*) and *b*) aforementioned are still present. Indeed, the demand flow of the MSP is identical to the demand flow of our initial production line and the characterization of its properties have to be discussed in order to look for optimal scheduling (case *a*)). Therefore, in the present thesis we shall start by assuming that system identification is already performed. Hence, we assume that the probabilistic laws characterizing the uncertainties are determined. The attention will then be focused on the dynamic scheduling of the MSP which minimizes the backorder and holding costs of the multiple finished goods surpluses and also minimizes the setup costs and/or time delays incurred when switching from one production type to another (case *b*). Let us first consider simple scheduling rules solving the MSP and discuss its efficiency:

i) **The Polling Policy:** The flexible facility first produce a fixed number P_j of one type of good then switch its production to another type, produce a fixed number P_k of this type, and so on until every type of good has been produced, then it restarts cyclically to the first produced type of good. The values P_j, $j \in \{1, 2, \ldots, N\}$ are chosen *a priory* at every cycle start.

ii) **The Minimal Stock Policy:** The flexible facility always produces the variant which has the smallest number of items in its inventory (this value may be negative when no goods are in the inventory and there are customers waiting to be served). When all the finished goods inventories are filled up, the facility is idle until a demand comes.

iii) **The Two Thresholds Policy:** We fix a minimal threshold d_j and a maximal threshold D_j for the finished good inventory of each product type $j \in \{1, 2, \ldots, N\}$. The minimal and maximal value may be different for each j. When one of the inventories falls below d_j we switch the production facility to this type of good and produce it until its stock reaches the level D_j. Then we look, among the other types, for another having its inventory level below its minimal threshold. If one exists, we switch the production to this type (say type k) and we produce it until its inventory level is D_k (if several exist we switch to the one having the lowest inventory). If none exists we carry on producing the type j until either one type k finds its stock level falling below d_k or the inventory of type j is filled up, then we let the facility be idle.

What is the efficiency of these heuristics?

i) The polling policy is a static policy in the sense that it does not react to environmental events. Indeed, once a cycle is started, no decisions are taken until the next one. This could produce truly poor results if, for example, the cycle time were long and suddenly demand rose higher than expected for a particular good. In this situation the customer for those goods would have to wait a long time before being served, which could considerably reduce customers satisfaction (this is really costly in a competitive environment).

ii) The minimal stock policy has catastrophic results when setup penalties are to be taken into account. Indeed, assuming that all inventories full to approximately the same level, then this policy will command to switch the production very often in a short interval of time. Because of the setup costs, this cannot possibly be optimal.

iii) The two thresholds policy precludes switching all too often from one type of good to another. It is therefore well adapted for production problems with setup penalties. Note however that in order to implement it in an actual factory, a difficult problem is

to find the adequate threshold d_j and D_j. Moreover these thresholds are fixed once an for all and do not depend on environmental events. This lack of reactivity certainly prevents the policy from being optimal and it is indeed what will be observed in part III below.

Being award that simple heuristics can not possibly lead to the optimal policy, we will study two mathematical formalisms affording to model the MSP and based on this model we will propose more efficient heuristics.

1.2 Mathematical Formalization of Decision-Making Problem

The very precursor to use mathematical formalizations for decision-making problems, is probably Blaise Pascal who, in the 17th century, proved the importance of believing in God by constructing his famous "wager" (see [39]):

[...] "*God is, or He is not. But to which side shall we incline? Reason can decide nothing here. There is an infinite chaos which separated us. A game is being played at the extremity of this infinite distance where heads or tails will turn up... Which will you choose then? Let us see. Since you must choose, let us see which interests you least. You have two things to lose, the true and the good; and two things to stake, your reason and your will, you knowledge and your happiness; and your nature has two things to shun, error and misery. Your reason is no more shocked in choosing one rather than the other, since you must of necessity choose... But your happiness? Let us weigh the gain and the loss in wagering that God is... If you gain, you gain all; if you lose, you lose nothing. Wager, then, without hesitation that He is.*"

Pascal, Pensées, 233.

His argument can be summarized as follows:

Available Choices	God exists	God does not exist
I believe	Reward $= C_{11}$	Loss $= C_{12}$
I do not believe	Loss $= C_{21}$	Reward $= C_{22}$

Here above I have tried to express Pascal's word in table format that should be read as follows: Available choices are in column 1, the assumption that God does actually exist in column2 and that he does not in column 3. The rows represent an individual's choice. He may believe (row 1) of not believe (row 2). The point of the exercise is to try a measure (albeit in a hypothetical manner) whether there is more to be gained and less to be lost from being a believer as opposed to a non-believer. We may gain this info from looking up the permuted reward and losses from the table. in symbols we have that C_{ij}, $i, j = 1, 2$ are the expected gain or lost obtained when:

- C_{11} we belive in God and He exists,
- C_{12} we believe in God and He does not exist.
- C_{21} we do not believe in God and He exists,
- C_{22} we do not believe in God and He does not exist.

Since this thesis is all about optimal decision taking under uncertainty (i.e. in a random environment) we will be dealing extensively with probabilities. Here is an example of

how we would use a simple probability measurement to the Pascal problem cited above. Assume that the probability of the existence of God is P_1 and the probability of his non-existence is P_2. Hence, the expected reward obtained by someone who believes in God is

$$S_1 = P_1 \times C_{11} + P_2 \times C_{12}$$

i.e. he will receive reward C_{11} with probability P_1 and reward C_{12} with probability P_2. Similarly, the expected reward for someone who does not believe in God is

$$S_2 = P_1 \times C_{21} + P_2 \times C_{22}.$$

In his "*Pensées*" [39], Pascal shows that S_1 is much larger than S_2 (i.e. $S_1 \gg S_2$). Indeed, assume that $P_1 = P_2$ (i.e. the probability of existence of God is equal to the probability of his non-existence). Now, remark that C_{11} has, for the believer, an infinite positive value (he will reach eternal life) and C_{12} has a finite positive value (his belief helps him to overcome his fear). On the contrary, C_{21} has an infinite negative value and C_{22} a finite negative value. It is therefore imperative to believe in God as S_1 brings infinite satisfaction while S_2 brings infinite desolation.

[...] "*But there is an eternity of life and happiness. And this being so, if there were an infinity of chances, of which one only would be for you, you would still be right in wagering one to win two, and you would act stupidly, being obliged to play, by refusing to stake one life against three at a game in which out of an infinity of chances there is one for you, if there were an infinity of an infinitely happy life to gain. But there is here an infinity of an infinitely happy life to gain, a chance of gain against a finite number of chances of loss, and what you stake is finite. It is all divided; wherever the infinite is and there is not an infinity of chances of loss against that of gain, there is no time to hesitate, you must give all...*"

Pascal, Pensées, 233.

While Pascal applied this mathematic formalization to a fundamental theological problem, the formalization of decision mechanisms went on to some more frivolous problems such as the card games, or random games, which were popular in the 18th century (examples are the contributions of Georges-Louis Buffon and the Bernoulli brothers, particularly Daniel). As emphasis in section 1.1, in this thesis we will apply the mathematical formalization to manufacturing problems.

Let us now introduce the mathematical notations needed in the sequel by modeling a general decision-making problem in a random environment: Consider a dynamical system consisting of N projects evolving with time. A Decision Maker (DM) engages these projects in parallel and we suppose that he can engage only $M < N$ of them at the same time (we usually say that the DM has a "limited capacity"). When a project is engaged by the DM we say that it is in its "active phase" otherwise we say that it is in its "passive phase". Let us write

$$X_j^{\pi_j}(t),\ j \in \{1, 2, \ldots, N\}$$

for the state of the project j at time t. Here, the superscript $\pi_j(t) \in \{a, p\}$ correspond to the operating state (active or passive) of project j at time t. Particularly $\pi_j = a$ when the DM engages project j at time t and $\pi_j = p$ otherwise. Assume that $\forall\, t \in \mathbb{R}_+$ we have:

$$X_j^{\pi_j}(t) \in \mathcal{X}_j.$$

We then say that $\mathcal{X}_j$ is the set of possible states of project j. We will use the notation

$$\vec{X}^{\vec{\pi}}(t) = (X_1^{\pi_1}(t), X_2^{\pi_2}(t), \ldots, X_N^{\pi_N}(t))$$

to characterize the global state of the system at time t. Note that the limited capacity of the DM implies that at most M of the N components of the vector $\vec{\pi}(t)$ have their value equal to "a". For example, suppose that $N = 3$ and $M = 1$, then:

$$\vec{X}^{\vec{\pi}}(t) = (X_1^p(t), X_2^a(t), X_3^p(t))$$

means that at time t, the project $j = 2$ is engaged and the projects $j = 1$, and $j = 3$ are disengaged. The evolution of all the projects will generally follow stochastic (i.e. random) processes and we impose an important concept, the Markov property of the processes:

Definition (Markov Process [41]) We say that a process is Markovian if its future probabilities are only determined by its most recent values.

Therefore, under the Markov assumption, the information contained in $\vec{X}^{\vec{\pi}}(t)$ at time t (i.e. the present state of the system) is sufficient in order to completely characterize the stochastic evolution of $\vec{X}^{\vec{\pi}}(\tau)$ for $\tau > t$. Let us now assume that we can construct "utility functions"

$$h_j^{\pi_j}(x_j),\ j \in \{1, 2, \ldots, N\}$$

which give the instantaneous reward gained when the project j is in state $X_j^{\pi_j}(t) = x_j$ (generally, $h_j^a(x_j) \neq h_j^p(x_j)$). Then, based on $\vec{X}^{\vec{\pi}}(t)$ and $h_j^{\pi_j}(X_j^{\pi_j}(t))$, the DM decides which project to engage with respect to his limited capacity and in order to maximize his global reward. Note that all the decisions taken by the DM at time t are summarized in the vector $\vec{\pi}(t)$ which assigns values "a" to each activated project and "p" to the others. The function

$$\vec{\pi}(t) : \{\mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_N\} \mapsto \{a, p\} \times \ldots \times \{a, p\}$$

$$(x_1, x_2, \ldots, x_N) \quad \mapsto \quad \vec{\pi}(t) \in \{a, p\}^N$$

will be called a "scheduling policy" and for ease of notation we will omit the arrow when no confusion can arise (i.e. $\vec{\pi}(t)$ will be written as $\pi(t)$ in the following). When the scheduling policy $\pi(t)$ makes his decisions based on the evolution of the system $\vec{X}^{\vec{\pi}}(t)$ we say that the policy is a "dynamic scheduling policy". Note, however, that not all policies $\pi(t)$ are of dynamic type. For example, when $M = 1$, we can construct a policy which engages each project j periodically during a fixed time T_j.

Let us now consider a time limit H (also called time horizon) during which we engage the projects following a scheduling policy $\pi(t)$. We can then define the global performance of the policy as

$$J_\pi(\vec{\pi}_0, \vec{x}_0) = \mathbf{E}_\pi \left\{ \int_0^H e^{-\beta s} \left[\sum_{k=1}^N h_k^\pi(X_k^\pi(s)) \right] ds - \sum_{\tau_i \leq H} e^{-\beta \tau_i} C_{\tau_i} \right\} \tag{1.1}$$

where

- $\vec{\pi}_0$ describes which are the M projects initially engaged.
- The τ_i are the instances at which it is decided to disengage a project and engage another one (i.e. the switching times).
- The C_{τ_i} are the switching costs incurred during the setups.
- The term $\mathbf{E}_\pi$ represents the expectation operator which computes the mean reward for all the possible realisations of the random process $\vec{X}(t)$. Note that when the dynamics of the system is deterministic, the presence of this operator is superfluous as $X_j(t)$ takes a single realization.
- The term $e^{-\beta t}$, with $\beta > 0$ is a discount factor which takes into account the fact that the future rewards are less attractive than ones available immediately. Specifically, one reward unit will correspond to $e^{-\beta t} < 1$ when received t units of time in the future.

Therefore, for a large value of β all the rewards gained in the future are almost negligible (the value of $e^{-\beta t}$ is close to 0) and maximizing the global satisfaction is equivalent to maximizing the immediate one (myopic policy). On the contrary, when β is small, future rewards are not negligible, thus, the myopic policy is not necessarily optimal.

Clearly, the first term of equation (1.1) corresponds to the sum of all discounted rewards received in each project j over the horizon H and the second term represents the sum of all discounted switching costs. Now the goal for the DM is to find the scheduling policy π^* which maximizes $J_\pi(\vec{\pi}_0, \vec{x}_0)$, i.e.

$$J_{\pi^*}(\vec{\pi}_0, \vec{x}_0) = \sup_\pi J_\pi(\vec{\pi}_0, \vec{x}_0).$$

Such a policy is called optimal.

2

Thesis Organization and New Results

The present thesis belong to the general research program of the Group QuaDeStra (Quality decision Strategy "http://ipr.epfl.ch") in which we view the supply-production-customer chain as being a far from thermodynamical equilibrium process [Hong 94] and [Schw. 97]. This point of view is summarized in the Figure 2.1 below where the position the present thesis (i.e. the dynamic scheduling problem) is located in the hatched part.

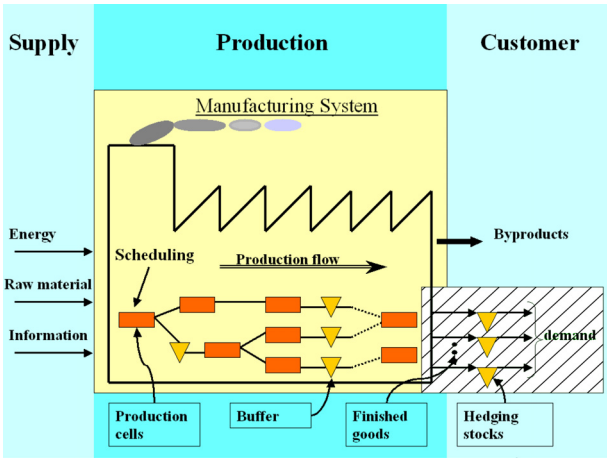

Fig. 2.1. Schematic view of a supply-production-customer chain. The domain of relevance of the present thesis is located in the highlighted green part of the production process.

In order to model our scheduling problem we will focus our attention on two classes of decision problems namely the "Multi-Armed Bandit Problem" (MABP) in part I and the "Restless Bandit Problem" (RBP) in part II. Because of the importance of setup penalties when switching production in real industrial environments, we generalize the MABP in part III by including switching costs an/or time delays. All the results obtained from these formalisms are applied to the flexible manufacturing problem in part IV. Let us just mention here the particularities of the MABP (with and without switching penalties) and the RBP:

a) **Multi-Armed Bandit Problem (without switching penality):**
   The following simplifications are included in the model:
   i) There is no switching cost.
   ii) The disengaged projects do not evolve in time (i.e. $X_j^p(t) \equiv X_j^p(t+dt)$)
   iii) The disengaged projects bring no reward (i.e. $h_j^p(x) \equiv 0$).
   iv) The DM can engage only one project at a time (i.e. $M = 1$).

v) The projects are independent (i.e. the evolution of project j does not influence the evolution of the other projects $k \neq j$).

b) **Multi-Armed Bandit Problem with switching penalties:**
The following simplifications are applied:
i) There are switching penalties each time we stop a project and we engage another one.
ii) Points ii) to v) are identical to the ones for MABP without switching penalty.

c) **Restless Bandit Problem:**
The following simplifications are included in the model:
i) The first point is similar to the MABP, i.e. there is no switching cost.
ii) The disengaged projects evolve in time and their evolution may be different when engaged or disengaged.
iii) The disengaged projects bring rewards.
iv) The DM can engage more than one project at a time (i.e. $1 \leq M < N$).
v) As in the MABP, the projects are independent (i.e. the evolution of project j does not influence the evolution of the other projects $k \neq j$).

Multi-Armed Bandits problems are in fact specific sub-problems of Restless Bandits problems. Therefore we can build a hierarchy of bandit problems starting with the simplest sub-problem (the Deteriorating MABP) to the more complex one (the RBP). This is resumed in the single table which follows:

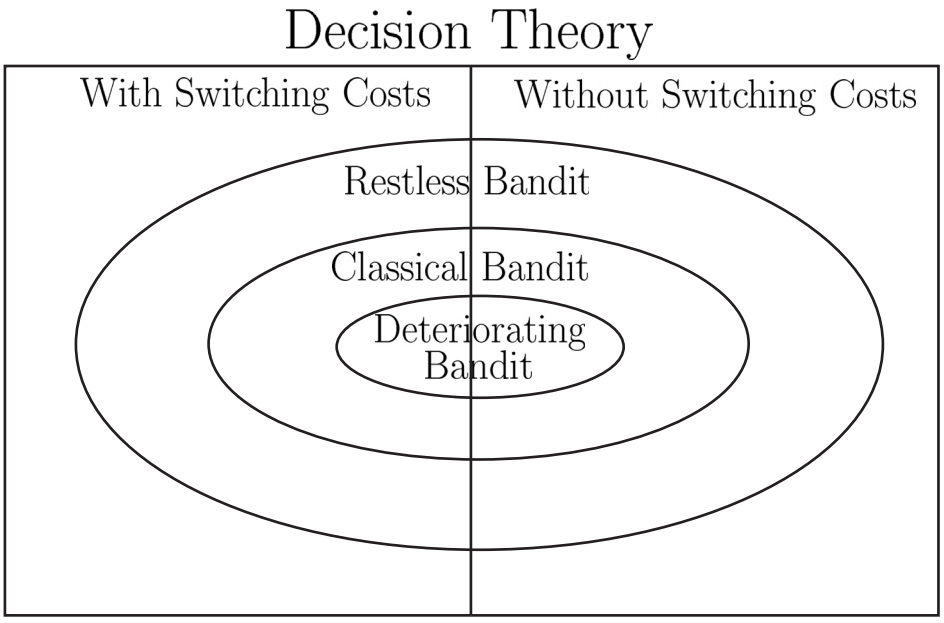

Fig. 2.2. Classification of Multi-Armed Bandits problems and Restless Bandits problems.

	MABP		RBP	
	Deterministic	Random	Deterministic	Random
Without switching penalties	1	2	3	4
With switching costs	5	6	7	8
With switching time	9	10	11	12
With switching costs and switching time	13	14	15	16

Our contribution spreads over classes 1 to 6 plus 9, 10, 13 and 14. Note that every section which presents a new result has its title beginning with a star $\left(\circledast\right)$. In particular:

- We compute explicitly the form of the Gittins indices when the evolution is given by a piecewise deterministic process which is intrinsically non-Markovian. This is the first class of non-Markovian examples in the literature for which the Gittins indices can be explicitly computed (see [31]).

- We studied several underlying random dynamics relevant to the industrial engineering context, (e.g. diffusion processes as well as birth and death processes). We obtained explicit generalized priority indices for RBP and the resulting dynamic scheduling was compared with exact results derived numerically by A. Ha [22]. Finally, using these results, we propose a sub-optimal heuristic solving the multi-items make-to-stock production problem (see [12] and [11])

- For the MABP with switching penalties, significant progress has been made by constructing a new class of MABP with setups for which the optimal policy can be explicitly constructed by recursion. Using this optimal derivation, we then propose a heuristic, approaching the optimal policy for general MABP with switching penalties (see [13]).

In short, the structure the thesis is as follows:

Each part begins with an introduction summarizing its content and ends with a small conclusion. We formally define the MABP and give some characteristic illustrations in chapter 3. In chapter 4 and appendix A, we review the optimal scheduling policy for MABP. We briefly present the elegant example given by Karatzas [24] in section 5.2, who computes explicitly the priority index for a MABP in continuous time, when the dynamics are driven by diffusion processes. We explicitly derived in section 5.3 the priority index function for MABP with a dynamics driven by piecewise deterministic evolution processes.

Note that, for the RBP, the optimal policy is not known, given general underlying dynamics and reward functions. However, a heuristic based on the priority index, known under the name "Whittle Relaxation", yields very efficient scheduling rules. The definition of the RBP can be read in chapter 7 and the derivation of the Whittle heuristic is described in chapter 8. Explicit computations of the indices are performed for different underlying dynamics:

- in sections 8.1.1 for a simple deterministic dynamics,
- in sections 8.1.2 for a diffusive dynamics,
- in sections 8.1.3 for a continuous time Markov chain dynamics.

The efficiency of these index policies is finally tested in sections 17.2.2 and 17.3 where we compare them to numerically derived optimal policies for a flexible machine able to produce two different product types.

For the MABP with switching penalties, it has been established that no policy based on priority indices can possibly be optimal (see section 12.1). Nevertheless index policies are attractive as they are simple to use. Therefore, in chapter 12, we construct a possible generalization of the Priority Index Policy based on two index functions for each project $j \in \{1, 2, \ldots, N\}$. Then, in chapter 14, we derive the optimal policy for a simple class of MABP with switching costs and, based on this solution, we show in chapter 15 that the proposed generalization of the Priority Index Policy leads to very interesting heuristics as it reproduces the generic form of the optimal policy. Even if the optimal policy for

MABP with switching penalties is generally not known, we can maintain that only dynamic scheduling will lead to an optimal policy. This is precisely what is observed in the example studied in chapter 14.

Finally, all the above result are used in part IV to discuss the scheduling policy of a flexible manufacturing production facility.

Part I

Multi-Armed Bandit Problem (MABP)

Introduction

The Multi-Armed Bandit Problem (MABP) is an idealized mathematical description of the conflict existing in the decision-making problems. The term "Multi-Armed Bandit Problem" comes from the case where the decisions to be taken are engagement of slot machines. In the vast domain of decision problems, the class of MABP certainly plays an important role as it is a building block used to model decision problems and it can be solved optimally. Indeed, in 1974, Gittins proved that the optimal policy for the MABP may be obtained by solving a family of simple stopping problems that associate with each project an urgency function (known as the Gittins index) in terms of which the optimal scheduling reads:

"At each instant produce the item exhibiting the highest priority index".

We briefly recall the definition of the MABP in chapter 3. Then, in chapter 4 we present the priority index policy based on the Gittins Index, which is the optimal policy for MABP (see appendix A). The remaining difficulty is to explicitly calculate the Gittins Index for given situations. As noted in [21], it is rather exceptional that the Gittins Index can be solved explicitly and only few examples are available in the literature among which we can cite:

- The simple deterministic example derived by Walrand [46] which is exposed in chapter 5.
- The class of deteriorating MABP, first proposed by Whittle [49], which is defined and solved in section 5.1.
- The MABP given by Karatzas [24], for with the underlying random dynamics is modeled by diffusion processes (see section 5.2)

We complete this list by proposing an original contribution, exposed in section 5.3, where we solve the class of alternating Markovian renewal processes (i.e. piecewise deterministic (PD) evolution processes). One motivation to study PD processes originates from the fact that they arise naturally in the fluid modeling of production flows delivered by failure-prone machines, [6], [17]. Finally we apply our results to a simple production scheduling problem in section 5.3.3.

Illustration of Multi-Armed Bandit Problems

The Gambling Machine Problem [49]: A natural illustration of MABP is the slot machine problem as follows: Suppose that a player faces two gambling machines. Assume that he knows that the first machine always distribute good rewards and assume that he knows nothing about the reward sequence of the second machine. From the point of view of immediately expected rewards, it would be preferable to use the first machine. However, if long-term performances are important, then it may be preferable to test the second machine, which has a chance of being better than the first one.

The Gold Mine Problem [18]: Consider the situation where a miner extracts gold from N different mines but can work only in one of them at a time. Every morning he has a choice of either continuing to work in the same mine as in the previous day or to change to another mine. He takes his decision accordingly to what he gained in the mines in the past. His goal is to maximize the extracted quantity of gold.

The Clinical Trials Problem [50]: A doctor has N alternative medical treatments to treat a particular illness. He can only use one of these treatments on each patient. Each time he uses a treatment he learns more about its efficiency. His goal is to find the treatment which is the best adapted for the illness. It is important to note that ethics play an important role in the present decision problem. Indeed, the doctor has to choose if he wants to maximize a short-term or a long term-reward as follows: To go for the short-term reward is to maximize the probability of success for the present patient. To go for the long-term reward means experimenting until he is confident to have found the best treatment.

Remark: An important common feature of all the above illustrations is the possibility for the Decision Maker (DM) to have a choice between several attitudes (e.g. which project to engage) at each moment in time. Another one is the impossibility for the DM to collect the reward of all the projects at the same time (we say that the DM has a limited capacity).

3

Definition of the Multi-Armed Bandit Problem

3.1 Discrete Time

The Multi-Armed Bandit Problem (MABP) exemplify the conflict of a decision maker (DM) having to choose at each instant of time one among N projects (also named arms from now on) as follows: Consider N dynamical projects. Let us assume that the DM may choose to continue with the currently engaged project or switch to another at each decision time t_i. Without loss of generality, we can assume that the DM makes a decision at each unit of time (i.e. $t_i = i \in \mathbb{N}$). Let us write $X_j(t_i)$ for the state of the project j at time t_i, then:

$$X_j(t_i) = x_j \in \mathcal{X}_j,$$

where $\mathcal{X}_j \subset \mathbb{R}^n$, $n > 0$ is the set of all possible states for the project j. If at time t_i, the decision is to engage the project j in the state x_j, then one gets an instant reward $h_j(x_j)$, where

$$h_j \ : \ \mathcal{X}_j \to \mathbb{R}$$

is a $\mathcal{C}^2$ uniformly bounded function. The rewards are discounted over time by a factor $0 < \beta < 1$. This means that the present value of one unit of gain equals β^t when received t units of time in the future. Each time the project j is engaged it moves into the state $X_j(t_{i+1})$, where $X_j(t_i)$ is a stochastic process. We assume the stochastic independence of the $X_k(t_i),\ k \in \{1, 2, \ldots, N\}$. Moreover, the stochastic processes $X_k(t_i)$ are assumed to be Markovian and stationary. Finally, we impose that the states of the $N-1$ disengaged projects remain frozen (i.e. $X_k(t_{i+1}) = X_k(t_i),\ \forall k \neq j$) and yield no reward. Note that, with these assumptions, the MABP belongs to the class of stationary Markov decision processes [41].

We write

$$\vec{X}(t_i) = (X_1(t_i), \ldots, X_N(t_i))$$

for the state of the N-armed MABP at time t_i. By abuse of notation we will also use $\vec{X}(t_i)$ instead of $(\vec{X}, \vec{h})$ to represent the N-armed Bandit problem itself. The decision to engage (or disengage) a project, as a function of time, is called a scheduling policy π. Therefore, at each decision time t_i a policy commands to engage one of the N projects (i.e. $\pi(t_i) \in \{1, 2, \ldots, N\}$). In the following, we will restrict our attention to the class of admissible policies defined by:

Definition 3.1 (Admissible policy [50]). *A policy π is admissible if $\pi(t_i)$ is a deterministic function depending only on the instantaneous state $X_j(t_i)$ of project j at time t_i (i.e. it is non-anticipating). In the following the set of all admissible policies is called $\mathcal{U}$.*

3 Definition of the Multi-Armed Bandit Problem

Let us define the indicator $I_j^\pi(t_i)$:

$$I_j^\pi(t_i) = \begin{cases} 1 & \text{if project } j \text{ is engaged at time } t_i \text{ under policy } \pi, \\ 0 & \text{otherwise} \end{cases}$$

Given a policy π, an initial condition $\vec{X}(t_0)$ and a discounting factor $0 < \beta < 1$, we can write the total discounted reward gained when engaging the MABP under policy π as

$$J^\pi(\vec{X}(t_0)) = E_\pi \left\{ \sum_{i=0}^{\infty} \beta^{t_i} \sum_{j=1}^{N} h_j(X_j(t_i)) I_j^\pi(t_i) \;\middle|\; \vec{X}(t_0) \right\}, \tag{3.1}$$

where $E_\pi\{ \cdot \mid \vec{X}(t_0)\}$ is the mathematical expectation under policy π conditional to the initial condition $\vec{X}(t_0)$. In equation (3.1) the first sum extends to all the decision times $t_i,\ i = 1, 2, \ldots$ and the second sum gives the reward of the engaged project at time t_i (when the engaged project is project j, then $I_j^\pi(t_i) = 1$ and $I_k^\pi(t_i) = 0,\ k \neq j$). The aim is to find a scheduling policy π^* which maximizes the total discounted reward, i.e.

$$J^{\pi^*}(\vec{X}(t_0)) = \sup_{\pi \in \mathcal{U}} J^\pi(\vec{X}(t_0)). \tag{3.2}$$

Such a π^* policy will be called "optimal". Let us now define the one-step operators L_j which describes the short time evolution of the expected global reward when it is decided to engage project j at time t_i:

$$L_j J^\pi(\vec{X}(t_i)) = h_j(X_j(t_i)) + \beta E[J^\pi(\vec{X}(t_{i+1}) \mid I_j^\pi(t_i) = 1)], \tag{3.3}$$

where the expectation operator E computes the mean value of $\vec{X}(t_{i+1})$ at time t_{i+1}. In term of the L_j operators $j \in \{1, 2, \ldots, N\}$, $J^{\pi^*}(\vec{X}(t_0))$ may be equivalently defined by using the Dynamic Programming (see [49] and [51]), as being the unique solution of:

$$J^{\pi^*}(\vec{X}(t_0)) = \max_{j=1,\ldots,N} L_j J^{\pi^*}(\vec{X}(t_0)).$$

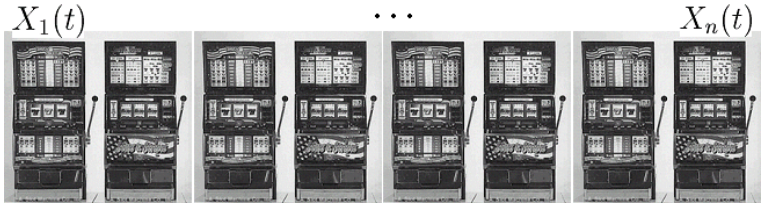

Fig. 3.1. The Gambling Machine Problem.

Remark: Applying this formalization to the illustrations given in the beginning of this section we have:

- For the Gambling Machine Problem (see Figure 3.1), we maximize the long-term performance when the discounting is not too steep (i.e. β is not too close to 0)
- For the Gold Mine Problem, the projects $X_j(t)$ are the quantity of gold extracted at time t in mine j and the discounting factor corresponds to the passive interest on the investment.

- For the Clinical Trials Problem the projects are the treatments, their state $X_j(t)$ is the degree of confidence in the efficiency of the treatment j and the discounting factor β is usually regarded as the ethical parameter. Indeed, as β varies from 0 to 1, one gradually changes from maximizing the short time reward (i.e. the treatment success of a single patient) to maximizing the long time reward (i.e. the research activity).

3.2 Continuous Time

In continuous time, the definition of the MABP is similar to the one in discrete time. The modifications to be included are the expected ones arising naturally from the continuous nature of the processes, namely:

- The $t_i,\ i \in \mathbb{N}$ denotes the sequence of increasingly ordered switching times occurring when it is decided to stop a project in order to engage another one.

- The discount factor is $e^{-\beta t}$.

Hence, equation (3.1) is rewritten as:

$$J^{\pi}(\vec{X}(t_0)) = E_{\pi}\left\{ \sum_{i=0}^{\infty} \int_{t_i}^{t_{i+1}} e^{-\beta t} \sum_{j=1}^{N} h_j(X_j(t)) I_j^{\pi}(t)\, dt \;\middle|\; \vec{X}(t_0) \right\}, \tag{3.4}$$

which is equivalent to

$$J^{\pi}(\vec{X}(t_0)) = E_{\pi}\left\{ \int_{0}^{\infty} e^{-\beta t} \sum_{j=1}^{N} h_j(X_j(t)) I_j^{\pi}(t)\, dt \;\middle|\; \vec{X}(t_0) \right\}. \tag{3.5}$$

Again the Dynamic Programming implies that the optimal cost function $J^{\pi^*}(\vec{X}(t_0))$ fulfills the following property:

$$\max_{j=1,\ldots,N} \left[h_j(X_j(t_0)) - \beta J^{\pi^*}(\vec{X}(t_0)) + L_j J^{\pi^*}(\vec{X}(t_0)) \right] = 0,$$

where L_j is the infinitesimal generator for $J^{\pi^*}(\vec{X}(t_i))$ (see [51] chapter 4, p.178).

4

Structure of the Optimal Policy for the MABP – The Gittins Index

The optimal policy for the MABP was first derived by J.C. Gittins [19] who introduced priority indices (Gittins indices) as follows:

Definition 4.1 (Priority Index Policy). *Given an N-armed MABP $\vec{X}(t_i)$, a Priority Index Policy is a scheduling policy, based on the existence of N stationary functions*

$$\begin{aligned} \nu g_j \ &: \mathcal{X}_j \to \mathbb{R} \qquad\qquad j = 1, \ldots, N \\ &\quad x_j \mapsto \nu g_j(x_j) \end{aligned}$$

depending only on the state of the project j alone (i.e. not depending on the other $N-1$ projects). In terms of the $\nu g_j(x_j)$, the Priority Index Policy reads: "At each decision time t_i engage the project exhibiting the largest index value $\nu g_j(X_j(t_i))$" (see Figure 4.1.

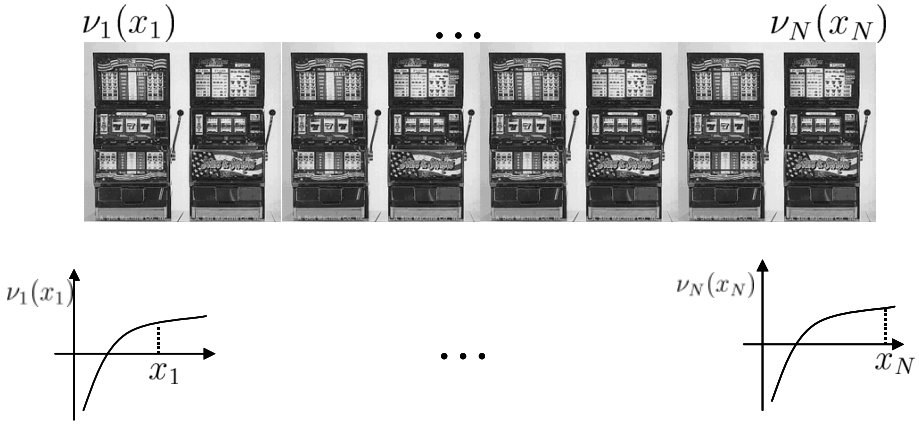

Fig. 4.1. Priority Index Policy.

4.1 The Gittins Index in Discrete Time

The computation of the priority indices is done by splitting the initial N-armed Bandit problem into N associated stopping problems (called as problems $\mathcal{SP}_j$, $j = 1, \ldots, N$):

Definition 4.2 (Problem $\mathcal{SP}_j$)). *Given an N-armed MABP $\vec{X}(t_i)$, the stopping problem $\mathcal{SP}_j$ associated with project $j = 1, \ldots, N$ are: Given a terminal reward*

$$\frac{\gamma}{1-\beta}$$

where $\gamma \in \mathbb{R}$ and $0 < \beta < 1$ is the discounting factor, find an optimal stopping time $\tau_j^*(\gamma) \in \mathbb{R}_+$ which maximizes the reward $J_j^\gamma(X_j(t_0))$ gained by engaging project j until time $\tau_j^*(\gamma)$, then stopping and collecting the terminal reward $\beta^{\tau^*(\gamma)}\frac{\gamma}{1-\beta}$.

For ease of notation we will omit the project index j in the optimal stopping time $\tau_j^*(\gamma)$ and write it simply as $\tau^*(\gamma)$. Given an initial condition $X_j(t_0)$, it can be proven (see [49]) that the maximal expected reward $J_j^\gamma(X_j(t_0))$ of the stopping problem $\mathcal{SP}_j$ is the unique bounded solution of the equation:

$$J_j^\gamma(X_j(t_0)) = \max\left[\frac{\gamma}{1-\beta}\ ;\ L_j J_j^\gamma(X_j(t_0))\right], \tag{4.1}$$

with the one-step operator L_j defined by equation (3.3). Indeed, the left alternative in the "max" of equation (4.1) correspond to stop while the right alternative is to continue. Intuitively, if the stopping reward γ increases, then the maximal expected reward $J_j^\gamma(X_j(t_0))$ should also increase. This intuition is correct and Whittle [49] proved that $J_j^\gamma(X_j(t_0))$ is a non-decreasing function of γ which equals $\frac{\gamma}{1-\beta}$ for γ sufficiently large.

Definition 4.3 (Gitting Index). *Given an N-armed MABP $\vec{X}(t_i)$, the Gittins index $\nu g_j(x_j)$ of project j for a given state $X_j(t_0) = x_j$, is defined as the smallest value of γ for which one has:*

$$J_j^\gamma(X_j(t_0)) = \frac{\gamma}{1-\beta}.$$

Remarks:

- By definition, the index $\nu g_j(X_j(t_0))$ is exactly the critical value of γ for which immediate stopping is optimal i.e.

$$L_j J_j^{\nu g_j(X_j(t_0))}(X_j(t_0)) = \frac{\nu g_j(X_j(t_0))}{1-\beta}. \tag{4.2}$$

- The index $\nu g_j(X_j(t_0))$ can be regarded as the minimal value of γ which **renders the following options equivalent:**

  a) Disengage project j immediately and collect the reward $\frac{\gamma}{1-\beta}$.

  b) Engage project j at least once, stop optimally thereafter at optimal stopping time $\tau^*(\gamma)$ and collect the reward $\beta^{\tau^*(\gamma)}\frac{\gamma}{1-\beta}$.

Definition 4.4 (Problem $\mathcal{P}_j$). *Given an N-armed MABP $\vec{X}(t_i)$, define N two-armed MABP (called problem $\mathcal{P}_j$, $j = 1, \ldots, N$) as follows:*

The first arm of the two-armed MABP $\mathcal{P}_j$ corresponds to the project j of $\vec{X}(t_i)$, with dynamics $X_j(t_i)$ and yielding a reward $h_j(X_j(t_i))$. The second arm of $\mathcal{P}_j$ (called project $\mathcal{T}$ from now on) has the "frozen" dynamics:

$$X_{\mathcal{T}}(t_i) \equiv \xi \in \mathcal{X}_{\mathcal{T}} \subset \mathbb{R},\ \forall\, t_i$$

and gives a constant reward γ when engaged (i.e. $h_{\mathcal{T}}(\xi) \equiv \gamma$, see Figure 4.2).

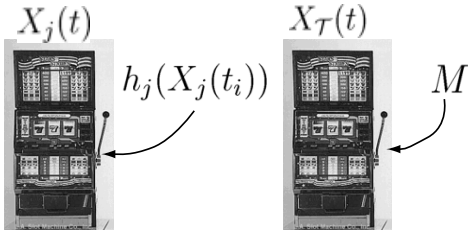

Fig. 4.2. Problem $\mathcal{P}_j$.

We will now prove that the problems $\mathcal{P}_j$ are equivalent to the problems $\mathcal{SP}_j$ in the sense that solving the problems $\mathcal{P}_j$ gives the stopping time $\tau_j^*(\gamma)$ and the reward $J_j^\gamma(X_j(t_0))$ of the problems $\mathcal{SP}_j$.

Lemma 4.5. *The problems $\mathcal{P}_j$ are stopping problems, equivalent to the problems $\mathcal{SP}_j$, $j = 1, \ldots, N$.*

Proof: Assume that initially project j is engaged. Then, once the project $\mathcal{T}$ is engaged, it will never be optimal to reengage the project j. Indeed, if at time t_i it is optimal to engage project $\mathcal{T}$, the states of the system at time t_{i+1} remain identical with those given at time t_i (the project j is "frozen"), hence by a recurrence argument engaging project $\mathcal{T}$ is optimal forever. From this observation, we conclude that problems $\mathcal{SP}_j$ and $\mathcal{P}_j$ are equivalent as they both engage project j until the stopping time $\tau_j^*(\gamma)$ which maximizes the global reward $J_j^\gamma(X_j(t_0))$.

□

Theorem 4.6 (Gittins Index). *Given the problem $\mathcal{P}_j$ associated with project j, the Gittins index for project j is ([18], [19], [46], [49]):*

$$\nu g_j(X_j(t_0)) = \sup_{\pi \in \mathcal{U}} \frac{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi - 1} \beta^{t_i} h_j(X_j(t_i))\right\}}{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi - 1} \beta^{t_i}\right\}}, \tag{4.3}$$

where $E_{x_j} = E_\pi[\ \cdot \mid X_j(t_0) = x_j]$, $\pi \in \mathcal{U}$ is an admissible (i.e. non anticipating) policy for problem $\mathcal{P}_j$ and t_{τ_π} is the stopping time at which project $\mathcal{T}$ is engaged under policy π (i.e. the numerator represents the expected reward of problem $\mathcal{P}_j$ until time t_{τ_π}).

Proof: For each admissible policy π for the problem $\mathcal{P}_j$, there is a corresponding stopping time t_{τ_π} at which it is required to engage project $\mathcal{T}$. As problem $\mathcal{P}_j$ and problem $\mathcal{SP}_j$ are equivalent, we have:

$$\tau^*(\gamma) = t_{\tau_{\pi^*}}$$

where $t_{\tau_{\pi^*}}$ is the optimal stopping time at which the optimal policy π^* (for problem $\mathcal{P}_j$) commands to engage project $\mathcal{T}$. Now, the minimal value of γ that makes equivalent to immediately disengage project j and collect the reward $\frac{\gamma}{1-\beta}$ or engage project j until time $t_{\tau_{\pi^*}}$, then stop and collect the reward $\beta^{t_{\tau_{\pi^*}}} \frac{\gamma}{1-\beta}$, is solution of:

$$E\left[\sum_{i=0}^{\tau_{\pi^*}-1} \beta^{t_i} h_j(X_j(t_i)) + \gamma \beta^{t_{\tau_{\pi^*}}}\right] = \gamma E\left[\sum_{i=0}^{\tau_{\pi^*}} \beta^{t_i}\right]. \tag{4.4}$$

Solving equation (4.4) for γ we get:

$$\gamma = \frac{E\left\{\sum_{i=0}^{\tau_{\pi^*}-1} \beta^{t_i} h_j(X_j(t_i))\right\}}{E\left\{\sum_{i=0}^{\tau_{\pi^*}-1} \beta^{t_i}\right\}}.$$

By definition, the Gittins index $\nu g_j(X(t_0))$ corresponds to this minimal value of γ. Hence, taking the "sup" over all admissible policies ensures that the summations in equation (4.3) are done with upper bound $i = \tau_\pi^* - 1$.

□

Theorem 4.7 (Optimality of the Gittins index). *The Priority Index Policy based on the Gittins index is the optimal policy for the MABP.*

Proof: See [19], [18], [46] or Appendix A.

□

Lemma 4.8. *The Gittins index of the "frozen" project $\mathcal{T}$ is*

$$\nu g_{\mathcal{T}}(X_{\mathcal{T}}(t_0)) = \gamma. \tag{4.5}$$

Proof: Consider problem $\mathcal{P}_{\mathcal{T}}$ associated with project $\mathcal{T}$ (i.e. a two-armed MABP with both projects having the "frozen" dynamics as defined for project $\mathcal{T}$). Suppose that the first project ($\mathcal{T}_1$) offers a constant reward of γ_1 and that the second project ($\mathcal{T}_2$) offers a constant reward of γ_2. Then the optimal policy of problem $\mathcal{P}_{\mathcal{T}}$ is to engage forever the project with the highest value of $\gamma_k,\ k = 1, 2$. This policy is achieved with priority indices $\nu g_{\mathcal{T}_k}(\xi)$ defined as:

$$\nu g_{\mathcal{T}_k}(\xi) = \gamma_k, \quad k = 1, 2.$$

□

Lemma 4.9. *The optimal stopping time $\tau^*(\gamma)$ of problem $\mathcal{SP}_j$ is:*

$$\tau^*(\gamma) = \inf\{t_i \mid \nu g_j(X_j(t_i)) \leq \gamma\}$$

Proof: Problem $\mathcal{SP}_j$ is equivalent to problem $\mathcal{P}_j$. As $\mathcal{P}_j$ is solved optimally with a Gittins index policy, the solution to the stopping problem $\mathcal{SP}_j$ is as follows: *"Engage project j as long as $\nu g_j(X_j(t_i)) > \nu g_{\mathcal{T}}(X_{\mathcal{T}}(t_i))$ otherwise stop and engage project $\mathcal{T}$ forever"*. From lemma 4.8 we know that $\nu g_{\mathcal{T}}(X_{\mathcal{T}}(t_i)) = \gamma$ which ends the demonstration.

□

Lemma 4.10. *When $h_j(x) \geq 0$, an optimal policy for a MABP commands never to be idle (see [49]).*

Proof: Assume *ad absurdum* that at time t_i the optimal policy commands to be idle during $T > 0$. Then, as the dynamics of the MABP is frozen during T, the state of the system is identical at time $t_i + T$ and at time t_i. Therefore, it is optimal to be idle at time $t_i + T$ and hence forever. This is clearly sub-optimal as $h_j(x) \geq 0$.

□

Remarks:

- The optimal policy for a two-armed MABP without switching cost is characterized by three subsets of $\mathcal{X}_1 \times \mathcal{X}_2$:

  $$S_1, S_2, S_{1\leftrightarrow 2} \subset \mathcal{X}_1 \times \mathcal{X}_2$$

  where

- The set S_1 contains the states $(x, y) \in \mathcal{X}_1 \times \mathcal{X}_2$ in which it is optimal to engage project 1:
$$S_1 = \{(x, y) \in \mathcal{X}_1 \times \mathcal{X}_2 \mid \nu g_1(x) > \nu g_2(y)\}.$$
- The set S_2 contains the states $(x, y) \in \mathcal{X}_1 \times \mathcal{X}_2$ in which it is optimal to engage project 2:
$$S_2 = \{(x, y) \in \mathcal{X}_1 \times \mathcal{X}_2 \mid \nu g_1(x) < \nu g_2(y)\}.$$
- The set $S_{1\leftrightarrow 2}$ contains the states $(x, y) \in \mathcal{X}_1 \times \mathcal{X}_2$ in which both previous decisions are equivalent:
$$S_{1\leftrightarrow 2} = \{(x, y) \in \mathcal{X}_1 \times \mathcal{X}_2 \mid \nu g_1(x) = \nu g_2(y)\}.$$

- When both projects are identical (same dynamics and reward function), the MABP is symmetric and we have that $\nu g_1(x) = \nu g_2(x)$, so
$$S_{1\leftrightarrow 2} = \{(x, y) \in \mathcal{X}_1 \times \mathcal{X}_2 \mid y = x\}$$
is a straight line.

4.2 The Gittins Index in Continuous Time

The continuous time version of section 4.1 is straightforward. Moreover, all the Theorems and Lemmas of the previous section hold in continuous time. We give here only the main changes:

- The terminal reward of the problem $\mathcal{SP}_j$ is $\frac{\gamma}{\beta}$.

- The maximal expected reward of the problem $\mathcal{SP}_j$ reads as:
$$J_j^{\gamma}(X_j(t_0)) = E\left\{\int_{t_0}^{\tau^*} e^{-\beta t} h_j(X_j(t))\, dt + e^{-\beta\tau^*}\frac{\gamma}{\beta} \;\middle|\; X_j(t_0)\right\}. \tag{4.6}$$

- **Definition 4.11 (Gittins index).** *Given the problem $\mathcal{P}_j$ associated with project j, the Gittins index for project j reads as ([18], [19], [46]):*
$$\nu g_j(X_j(t_0)) = \frac{E\left\{\int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\, dt\right\}}{E\left\{\int_{t_0}^{\tau^*} e^{-\beta t}\, dt\right\}}. \tag{4.7}$$

5

Examples of MABP and Their Explicit Solutions

In order to consolidate the notations and definitions presented in section 3.1, we reproduce here a simple deterministic MABP in discrete time proposed by Walrand in [46].

Consider three sequences of nonnegative numbers $X_j(t_i)$:

$$\begin{aligned} X_1(t_i) &= (3,2,4,1,0,0,\ldots), \\ X_2(t_i) &= (2,3,2,0,0,\ldots), \\ X_3(t_i) &= (2,1,4,0,0,\ldots) \end{aligned}$$

and assume that the discounting factor is $\beta = 0.5$. We define a MABP as follows: *"At each decision time t_i, choose one of the sequences, receive a reward corresponding to the first number of the chosen sequence and delete it from the sequence (e.g. if at time t_0 we choose the first sequence, we receive a reward of 3 and the first sequence becomes $X_1(t_{i+1}) = (2,4,1,0,\ldots)$). The goal is to maximize the global discounted reward $J^{\pi}(\vec{X}(t_0))$ given by equation (3.1)"*.

For project 1, we denote $\nu g_1(3)$ the value of the Gittins index when the initial condition is $X_1(t_0) = 3$ (i.e. we start the sequence of rewards at the beginning). Similarly, we denote $\nu g_1(2)$ the value of the Gittins index when the initial condition is $X_1(t_0) = 2$ (i.e. we already received the first reward of project 1 and start the sequence at the second position) and so on for each state of every project. Computing the Gittins index of project 1 we find:

$$\nu g_1(3) = \max\left\{3, \frac{3+2(0.5)}{1+0.5}, \frac{3+2(0.5)+4(0.5)^2}{1+0.5+(0.5)^2}, \frac{3+2(0.5)+4(0.5)^2+1(0.5)^3}{1+0.5+(0.5)^2+(0.5)^3}\right\} = 3$$

Similarly, we have

$$\nu g_1(2) = \tfrac{8}{3},\ \nu g_1(4) = 4,\ \nu g_1(1) = 1,\ \nu g_1(0) = 0,$$

$$\nu g_2(2) = \tfrac{7}{3},\ \nu g_2(3) = 3,\ \nu g_2(2) = 2,\ \nu g_2(0) = 0,$$

$$\nu g_3(2) = 2,\ \nu g_3(1) = 2,\ \nu g_3(4) = 4,\ \nu g_3(0) = 0,$$

The Priority Index Policy then corresponds to receiving the rewards of the three sequences in the following order:

$$1st,\ 1st,\ 1st,\ 2nd,\ 2nd,\ 3rd,\ 3rd,\ 3rd,\ 2nd,\ 1st,$$

i.e.

$$(3,2,4,2,3,2,1,4,2,1,0,\ldots),$$

which yields a global reward of

$$J^{\pi}(3,2,2) = 3 + \frac{1}{2}2 + \frac{1}{2}^{2}4 + \frac{1}{2}^{3}2 + \frac{1}{2}^{4}3 + \frac{1}{2}^{5}2 + \frac{1}{2}^{6}1 + \frac{1}{2}^{7}4 + \frac{1}{2}^{8}2 + \frac{1}{2}^{9}1 = 5.55664$$

Note that if we received the rewards as follows:

$$1st,\ 1st,\ 1st,\ 2nd,\ 2nd,\ 2nd,\ 3rd,\ 3dr,\ 3rd,\ 1st,$$

we would obtain the same global reward. This is because $\nu g_3(2) = \nu g_2(2)$. In fact, when at a decision time t_i, several projects have the same maximal index value we can engage anyone of them indifferently. We can easily check that 5.55664 is the optimal reward for this problem.

5.1 Deteriorating MABP

Let us introduce the class of "Deteriorating MABP". This is a primordial class of Bandit problems as every MABP can be reduced to a deteriorating MABP by a suitable construction [27] (see also chapter 13).

Definition 5.1 (Deteriorating MABP [49]). *We say that a MABP $\vec{X}(t_i)$ is deteriorating, if for all $j = 1, \ldots, N$, $J_j^{\gamma}(X_j(t_i))$ (defined in equation (4.1)) is non-increasing in t_i. We shall write DMABP for the class of deteriorating MABP.*

Theorem 5.2 (Whittle [49]). *The Gittins index for a DMAPB $X_j(t_i)$ with reward function $h_j(x)$ is:*

$$\nu g_j(x) = h_j(x). \tag{5.1}$$

Proof: Equation (4.1), written out explicitly is

$$J_j^{\gamma}(X_j(t_0)) = \max\left[\frac{\gamma}{1-\beta}\ ;\ h_j(X_j(t_0)) + \beta E[J_j^{\gamma}(X_j(t_1)) \mid X_j(t_0)]\right].$$

Now, letting $\gamma = \nu g_j(X_j(t_0))$, we have, by definition, that $J_j^{\gamma}(X_j(t_0)) = \frac{\gamma}{1-\beta}$ and stopping is optimal at $X_j(t_0)$. Thus, as the function $J_j^{\gamma}(X_j(t_i))$ is decreasing in t_i, stopping will certainly be optimal at $X_j(t_1)$ and

$$E[J_j^{\gamma}(X_j(t_1)) \mid X_j(t_0)] = \frac{\gamma}{1-\beta}.$$

Now, by definition of the Gittins index, we have that the decision to continue with project j or stop should give the same expected reward at position $X_j(t_0)$ (as $\gamma = \nu g_j(X_j(t_0))$). This yields the relation

$$\frac{\gamma}{1-\beta} = h_j(X_j(t_0)) + \beta\frac{\gamma}{1-\beta}.$$

Solving this equation for γ gives the equation (5.1). Accordingly, to choose the project with the maximal value of the index is to choose the project with the maximal immediate return.

□

Remark: The Priority Index Policy for a DMABP is a one-step look-ahead rule as it directly maximizes the instantaneous reward $h_j(X_j(t_0))$. Hence, we say that this policy is myopic.

Lemma 5.3 ([49]). *A MABP is a DMABP if and only if $h_j(X_j(t_i))$ is non-increasing in t_i, $\forall\ j = 1, \ldots, N$.*

Proof: If $h_j(X_j(t_i))$ is non-increasing in t_i, $\forall\ j = 1, \ldots, N$, we clearly have, by definition, that $J_j^\gamma(X_j(t_i))$ is non-increasing in t_i. To prove the converse, let us show that function $\nu g_j(X_j(t_i))$ is non-increasing. Then, as $\nu g_j(X_j(t_i)) = h_j(X_j(t_i))$ this will prove the non-increasing property of the reward function $h_j(X_j(t_i))$.

We want to show that $\nu g_j(X_j(t_1)) \leq \nu g_j(X_j(t_0))$. Assuming that $\gamma = \nu g_j(X_j(t_0))$; then, by definition of the reward function we have that

$$J_j^\gamma(X_j(t_0)) = \gamma.$$

As we deal with deteriorating MABP

$$J_j^\gamma(X_j(t_1)) \leq J_j^\gamma(X_j(t_0)).$$

Finally, by definition of $J_j^\gamma(X_j(t_0))$ this implies

$$\gamma \leq J_j^\gamma(X_j(t_1)) \leq J_j^\gamma(X_j(t_0)) = \gamma,$$

hence, stopping is optimal at $X_j(t_1)$. We end the demonstration using lemma 4.9 and we have that

$$\nu g_j(X_j(t_1)) \leq \gamma = \nu g_j(X_j(t_0)).$$

□

Illustration

Here is a simple example proposed by Whittle [49] to illustrate the class of DMABP.

Suppose there are N sites, at each of which there may be a treasure with probability P_j. Let v_j be the expected value of the treasure at site j and C_j the cost of examining the site for one unit of time. Supposed that there is a treasure at site j, then the probability that it will be found in a search over one unit of time is α_j. Hence, the expected immediate reward from an examination of site j reads as:

$$H_j = \alpha_j P_j v_j - C_j.$$

Let us then construct the following MABP:

- The projects $j = 1, 2, \ldots, N$, are the N sites.
- The projects states are $X_j(t_0) = P_j$ and $X_j(t_i) = \tilde{P}_j(t_i)$ where $\tilde{P}_j(t_i)$ is the posterior probability that there is a treasure at site j conditional to the fact that we didn't find it, although we looked for it at time t_{i-1}. Thus, when no treasure is found, by Bayes' theorem
$$X_j(t_i) = \tilde{P}_j(t_i) = \frac{\tilde{P}_j(t_{i-1})(1-\alpha_j)}{\tilde{P}_j(t_{i-1})(1-\alpha_j) + (1-\tilde{P}_j(t_{i-1}))}$$
and if the treasure is found, $\tilde{P}_j(t_i)$ increases to 1 but, since the treasure is then removed, $\tilde{P}_j(t_i)$ effectively falls to zero at the next step.
- The reward function is
$$h_j(x) = \alpha_j x v_j - C_j.$$

Note that with these definitions

$$h_j(X_j(t_i)) = \alpha_j X_j(t_i) v_j - C_j$$

is a non-increasing function (as $X_j(t_i)$ is non-increasing) and by Lemma 5.3, the defined MABP is a DMABP. Therefore, using Theorem 5.2, the optimal policy is to examine the site j for which $h_j(X_j(t_i)$ is maximal.

5.2 MABP with a Dynamics Driven by Diffusion Processes

In this section we briefly review the elegant example given by Karatzas [24] which will be needed in section 5.3. In [24], Karatzas computes explicitly the Gittins index for a MABP in continuous time, when the dynamics are driven by diffusion processes as follows:

$$dX_j(t) = \mu_j(X_j(t))dt + \sigma_j(X_j(t))dW(t),$$

$$X_j(t_0) = x_j^0,$$

with $dW(t)$ a White Gaussian noise process. The reward functions $h_j(x)$, $x \in \mathbb{R}$ are assumed to be strictly increasing with bounded, continuous first and second derivatives and they satisfy the conditions (see Figure 5.1

$$\lim_{x\to\infty} h_j(x) = K, \quad \lim_{x\to-\infty} h_j(x) = k, \quad \lim_{|x|\to\infty} h_j'(x) = 0.$$

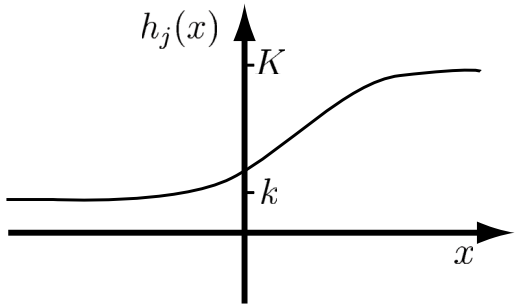

Fig. 5.1. Limit behaviours of $h_j(x)$.

For simplicity, we will derive here the Gittins index only for the particular case when the drift and the variance of the diffusion processes are constant i.e.

$$\mu_j(X_j(t)) \equiv \mu_j \quad \text{and} \quad \sigma_j(X_j(t)) \equiv \sigma_j.$$

To this end, following Whittle, we compute $J_j^\gamma(x_j^0)$, the maximal expected reward of the stopping problem $\mathcal{SP}_j$ (given by equation (4.6)) with initial condition $X_j(t_0) = x_j^0$. The increasing nature of the reward function $h_j(x)$ suggests that the continuation region for the stopping problem should be an open interval $]b_j(\gamma), \infty[$. Using the Dynamic Programing and the Itô's rule [25], it can be shown that in the continuation region, $J_j^\gamma(x_j^0)$ is solution of:

$$\frac{1}{2}\sigma_j^2 \frac{d^2}{dx^2} J_j^\gamma(x_j^0) + \mu_j \frac{d}{dx} J_j^\gamma(x_j^0) - \beta J_j^\gamma(x_j^0) + h_j(x_j^0) = 0. \tag{5.2}$$

Due to the linearity of equation (5.2), its general solution in $\mathbb{R}$ is:

$$J_j^\gamma(x_j^0) = A_j e^{-w_j^+ x_j^0} + B_j e^{w_j^- x_j^0} + S_j(x_j^0)$$

with the notations:

$$w_j^+ = \frac{\sqrt{\mu_j^2 + 2\beta\sigma_j^2} + \mu_j}{\sigma_j^2} > 0,$$

$$w_j^- = \frac{\sqrt{\mu_j^2 + 2\beta\sigma_j^2} - \mu_j}{\sigma_j^2} > 0$$

where A_j and B_j are integration constants and $S_j(x)$ are the particular solutions of equation (5.2) corresponding to engaging the project X_j forever with initial condition $X_j(t_0) = x$. We obtain:

$$S_j(x) = E_x \int_0^\infty e^{-\beta t} h_j(X_j(t))\, dt =$$

$$= \frac{2}{\sigma_j^2(w_j^+ + w_j^-)} \left[e^{-w_j^+ x} \int_{-\infty}^{x} h_j(y)\, e^{w_j^+ y} dy + e^{w_j^- x} \int_{x}^{\infty} h_j(y)\, e^{-w_j^- y} dy \right].$$

Note that when $x_j^0 = +\infty$ the optimal reward of the stopping problem is attained by engaging the project j forever. From the fact that $w_j^+ > 0$ and $w_j^- > 0$ we have:

$$\lim_{x\to\infty} J_j^\gamma(x) = \lim_{x\to\infty} S_j(x) \Rightarrow B_j = 0.$$

On the interval $]-\infty, b_j]$ the maximal reward of the stopping problem is attained by disengaging the project j immediately and receiving the stopping reward $\frac{\gamma}{\beta}$, i.e.

$$J_j^\gamma(x_j^0) = \frac{\gamma}{\beta}, \quad x_j^0 \in\,]-\infty, b_j].$$

By a smooth fitting argument (see chapter 8 for more details on the smooth fitting) we get:

$$\frac{d}{dx} J_j^\gamma(x)\Big|_{x=b_j(\gamma)} = 0. \tag{5.3}$$

Given γ, elimination of A_j in equation (5.3) yields the determining equation for the boundary value of the continuation region $b_j(\gamma) = b$. Conversely, given $b_j(\gamma) = \tilde{b} \in \mathbb{R}$, equation (5.3) enables to determine the value of γ such that the boundary value of the continuation region is exactly $\tilde{b}$. Karatzas shows that this γ is unique for each $\tilde{b}$. We can hence construct the function

$$\begin{aligned} \nu g_j \;:\; & \mathbb{R} \to \mathbb{R} \\ & x \mapsto \nu g_j(x) \end{aligned}$$

such that

$$b_j(\nu g_j(x)) = x.$$

It can be shown [24], that the function $\nu g_j(x)$ is non-decreasing. Therefore, given x_j^0, the value of γ solving equation (5.3) (i.e. $\gamma = \nu g_j(x_j^0)$) is the smallest value of γ for which it is optimal, for the stopping problem $J_j^\gamma(x_j^0)$, to immediately disengage project j when in position x_j^0, i.e.

$$J_j^{\nu g_j(x_j^0)}(x_j^0) = \frac{\nu g_j(x_j^0)}{\beta}.$$

Then, by definition, the function $\nu g_j(x)$ is the Gittins index of project j. After straightforward but lengthy algebra one obtains

$$\nu g_j(x) = \int_0^\infty h_j(x + \frac{y}{w_j^-}) e^{-y}\, dy \tag{5.4}$$

5.3 ⊛ MABP with a Dynamics Driven by Piecewise Deterministic Evolution Processes

In the beginning of this section we consider a single process X_j, we will therefore omit the item index j as long as no confusion arises. In order to compute the priority index $\nu g_j(x)$ associated with a class of piecewise deterministic processes, we will solve the stopping problem $\mathcal{P}_j$ of the type discussed in chapter 4.

5.3.1 ⊛ Optimal Stopping Problem

Here we focus our attention on the two-states random velocity model as follows: The process $X(t)$ evolves linearly in time with two possible velocities, velocity U or velocity D i.e.

$$X(t) = Ut + x_0 \quad \text{or} \quad X(t) = Dt + x_0.$$

Hence, its scalar time evolution for $t \geq 0$ is given by:

$$\frac{d}{dt}X(t) = I(t), \;\; X(0) = x, \;\; I(0) = \theta \in \{U, D\}, \;\; \text{and} \;\; X(t) \in \mathbb{R} \tag{5.5}$$

where $I(t)$ is an alternating Markovian renewal process ([42] and [40]) taking the alternate constant values U (up) and D (down). Assume that the sojourn times in the U and D phases are exponentially distributed with parameters λ and μ respectively. We further impose:

$$D\lambda + U\mu > 0 \;\; \text{with} \;\; D < 0 \;\; \text{and} \;\; D + U > 0$$

so that the expected value of $X(t)$ increases with time. Note that the path realizations of the stochastic differential equations equation (5.5) are continuous in contrast to the class of dynamics discussed in [9] where jumps in the realization are present. A typical realization of the solution of equation (5.5) is sketched in Figure 5.2. It is important to note that the process $X(t)$ solution of equation (5.5) is not Markovian. However the pair process defined by $\zeta(t) := (X(t), I(t)) \in \mathbb{R} \times \{D, U\}$ is itself a Markov process. Hence, the optimal reward equation (4.6) of the stopping problem SP must be defined with the variable $\zeta(t)$ and reads as:

$$J^{\gamma}(X(t_0), I(t_0)) = E\left\{ \int_{t_0}^{\tau^*} e^{-\beta t} h(X(t))\, dt + e^{-\beta \tau^*} \frac{\gamma}{\beta} \;\middle|\; X(t_0), I(t_0) \right\} \tag{5.6}$$

Let us assume that the reward function $h(x)$ is a twice differentiable function, which is strictly increasing. In addition, we follow [49] and [24] and impose the following asymptotic properties:

$$\lim_{x \to \infty} h(x) = K, \quad \lim_{x \to -\infty} h(x) = k, \quad 0 \leq k < K, \tag{5.7}$$

with $\beta > 0$ and:

$$\lim_{|x| \to \infty} \frac{d}{dx} h(x) = \lim_{|x| \to \infty} h'(x) = 0,$$

To solve the optimal stopping problem, we proceed as usual by applying a classical Dynamic Programming argument: *"When the process $(X(t), I(t))$ defined by equation (5.5) is in a given state, we can either wait a certain period of time before stopping or stop immediately"*. The resulting rewards are computed by assuming that the subsequent decisions are optimal. Accordingly, we can write:

i) If at time t, the decision is to wait for an (infinitesimal) time ξ before stopping, the optimal reward $J^{\gamma}(X(t), I(t))$ is greater than the running reward from t to $t + \xi$ and the optimal reward from time $t + \xi$ onwards. This yields the inequality:

$$J^{\gamma}(X(t), I(t)) \geq \int_t^{t+\xi} e^{-\beta(s-t)} h\left(X(s)\right) ds + e^{-\beta \xi} J^{\gamma}(X(t+\xi), I(t+\xi)). \tag{5.8}$$

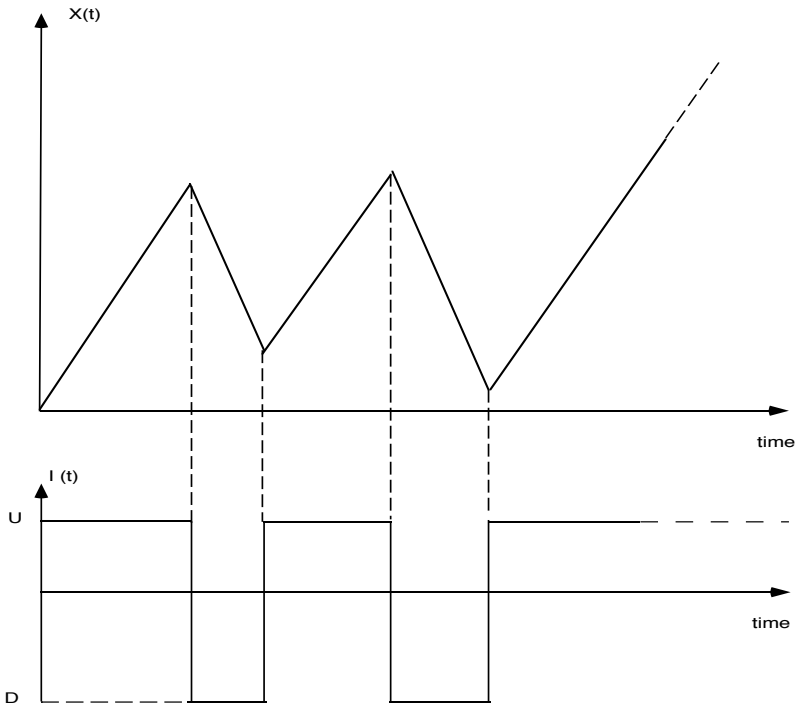

Fig. 5.2. Qualitative behavior of the solution of equation (5.5) as a function of a realization of the noise $I(t)$.

ii) If the decision at time t is to stop immediately, the reward is precisely:

$$J^\gamma(X(t), I(t)) = \frac{\gamma}{\beta}.$$

Let us focus on case i) for which we consider two possible rewards depending on the realization of the process $I(t)$, namely: $I = U$ or $I = D$. To these two possibilities, we associate the reward functions $J^\gamma(x, U)$ and $J^\gamma(x, D)$ respectively. With this notation and the Markov character of the alternating process $I(t)$, the first order time expansion of equation (5.8) yields:

$$0 \geq h(x) - (\beta + \lambda)J^\gamma(x, U) + U\frac{d}{dx}J^\gamma(x, U) + \lambda J^\gamma(x, D) \tag{5.9}$$

and

$$0 \geq h(x) - (\beta + \mu)J^\gamma(x, D) + D\frac{d}{dx}J^\gamma(x, D) + \mu J^\gamma(x, U). \tag{5.10}$$

Following the same line as Karatzas [24], we observe that the strictly increasing nature of $h(x)$ suggests that the stopping problem should have three regimes holding in three adjacent intervals $\mathcal{O}_j$, $j = 1, 2, 3$ defined by two thresholds $b_u < b_d$. In the interval $\mathcal{O}_1$ we continue to engage project j for both states U of D of the noise. In the interval $\mathcal{O}_2$ we continue to engage project j for the state U of the noise and disengage it when the

state is D. In the interval $\mathcal{O}_3$ we disengage project j for both the state U of D of the noise. Accordingly, we shall write:

a) $\mathcal{O}_1 = [b_d, \infty)$,

b) $\mathcal{O}_2 = [b_u, b_d]$,

c) $\mathcal{O}_3 = (-\infty, b_u]$.

Solving equations (5.9) and (5.10), we now construct the reward functions $J^{\gamma}_{(i)}(x, D)$ and $J^{\gamma}_{(i)}(x, U)$, $i = 1, 2, 3$, valid in the intervals $\mathcal{O}_1$, $\mathcal{O}_2$, and $\mathcal{O}_3$ respectively. The qualitative behavior of the reward functions is sketched in Figure 5.3.

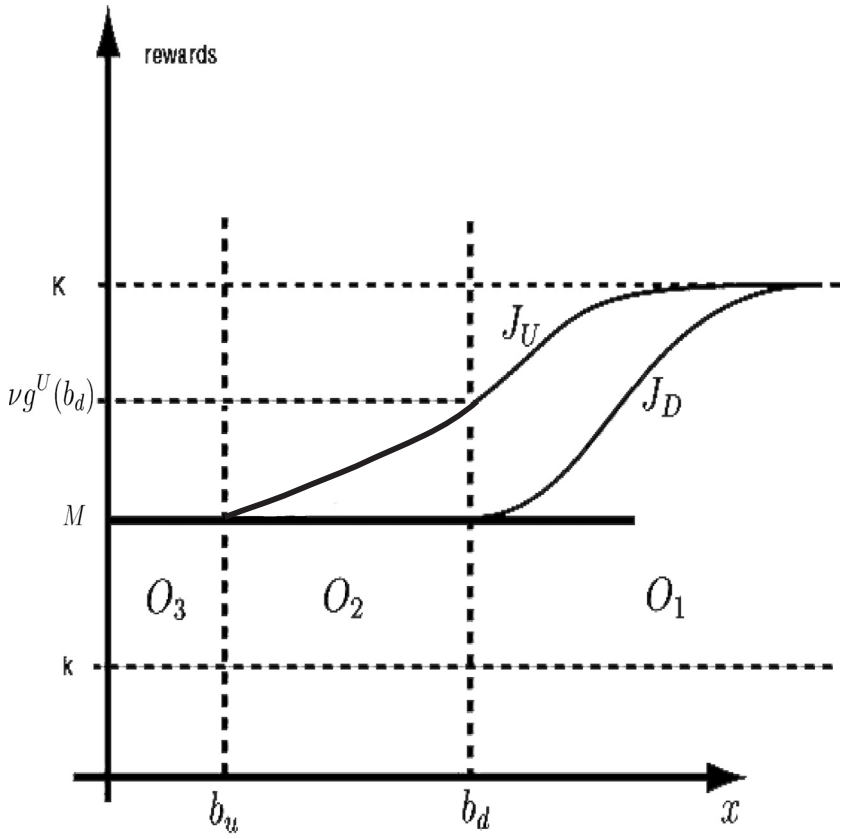

Fig. 5.3. Qualitative behavior of the rewards functions $J^{\gamma}(x, U)$ and $J^{\gamma}(x, D)$ in the intervals $\mathcal{O}_k$, $k = 1, 2, 3$.

a) **Solution in the interval** $\mathcal{O}_1$: This is a continuation region for both $J^{\gamma}_{(1)}(x, D)$ and $J^{\gamma}_{(1)}(x, U)$. In this region equations (5.9) and (5.10) hold simultaneously with equality. Before solving them, let us define:

$$\delta = D(\beta + \lambda) + U(\beta + \mu) > 0, \quad \text{and} \quad \rho = (\beta + \mu + \lambda)$$

and

$$h^D(x) = \rho h(x) - D h'(x) \quad \text{and} \quad h^U(x) = \rho h(x) - U h'(x).$$

Then, the solutions of the linear differential equations (5.9) and (5.10) are derived by determining the associate eigenvalues which read as:

$$-\gamma^+ = \frac{1}{2DU}\left[\delta + \sqrt{\delta^2 - 4DU\rho\beta}\right],$$

and

$$\gamma^- = \frac{1}{2DU}\left[\delta - \sqrt{\delta^2 - 4DU\rho\beta}\right].$$

As $\gamma^- > 0$, the solution attached to this eigenvalue does not converge for $x \to \infty$ and must then be rejected. Therefore, we only need to calculate the eigenvector associated with $-\gamma^+$ which read as $(\alpha, 1)$ with:

$$\alpha = \frac{-1}{2\mu U}\left(\delta - 2U(\beta+\mu) + \sqrt{\delta^2 - 4DU\rho\beta}\right).$$

Finally, the general solution of equations (5.9) and (5.10) read as:

$$J^{\gamma}_{(1)}(x, D) = A_{d,1}e^{-\gamma^+ x} + p_d(x), \quad \text{for} \quad x > b_d \tag{5.11}$$

and

$$J^{\gamma}_{(1)}(x, U) = \alpha A_{d,1}e^{-\gamma^+ x} + p_u(x), \quad \text{for} \quad x > b_d \tag{5.12}$$

with $A_{d,1}$ being an integration constant.

The functions $p_d(x)$ and $p_u(x)$ are particular solutions. They represent the rewards when running forever, (i.e. when no stopping occurs). Hence, from equation (5.7) we have that $p_u(x) > k$ and $p_d(x) > k$. In our case, two particular solutions of the inhomogeneous equations (5.9) and (5.10) respectively read as:

$$p_u(x) = E_{(x,U)}\int_0^{\infty} e^{-\beta t}h^D\left(X(t)\right)\,dt =$$

$$= \mathcal{N}\left[e^{-\gamma^+ x}\int_{-\infty}^{x} h^D(y)e^{\gamma^+ y}dy + e^{\tilde{\gamma}x}\int_x^{\infty} h^D(y)e^{-\tilde{\gamma}y}dy\right]$$

and

$$p_d(x) = E_{(x,D)}\int_0^{\infty} e^{-\beta t}h^U\left(X(t)\right)\,dt =$$

$$= \mathcal{N}\left[e^{-\gamma^+ x}\int_{-\infty}^{x} h^U(y)e^{\gamma^+ y}dy + e^{\tilde{\gamma}x}\int_x^{\infty} h^U(y)e^{-\tilde{\gamma}y}dy\right],$$

with:

$$\mathcal{N} = \left[-DU(\gamma^+ + \tilde{\gamma})\right]^{-1}$$

and

$$\tilde{\gamma} = \gamma^+ + \frac{\delta}{DU} \geq 0.$$

b) **Solution in the interval** $\mathcal{O}_2$: This is a stopping region when $I(t) = D$ and hence $J^{\gamma}_{(2)}(x, D) = \frac{\gamma}{\beta}$ and a continuation region when $I(t) = U$. Hence the use of equation (5.9) with $J^{\gamma}(x, D) = \frac{\gamma}{\beta}$ immediately yields:

$$J^{\gamma}_{(2)}(x, U) = e^{\omega x}\left[A_{u,2} - \int_0^x \left[\frac{\beta h(s) + \lambda\gamma}{\beta U}\right]e^{-\omega s}ds\right], \tag{5.13}$$

with $A_{u,2}$ being an integration constant and $\omega = \frac{\beta+\lambda}{U}$.

c) **Solution in the interval** $\mathcal{O}_3$: This is a global stopping region and we therefore have

$$J^{\gamma}_{(3)}(x,U) = J^{\gamma}_{(3)}(x,D) = \frac{\gamma}{\beta}$$

From equations (5.11), (5.12), (5.13), we see that we have to determine two constants of integration $A_{d,1}$ and $A_{u,2}$. Moreover the thresholds b_u and b_d are yet unknown. Hence, we need four relations to determine these four unknowns. These relations are (see figure (5.3)):

i) $J^{\gamma}_{(1)}(b_d,D) = \frac{\gamma}{\beta}$.

ii) $\frac{d}{dx}J^{\gamma}_{(1)}(x,D)\mid_{(x=b_d)} = 0$.

iii) $J^{\gamma}_{(2)}(b_u,U) = \frac{\gamma}{\beta}$.

iv) $J^{\gamma}_{(1)}(b_d,U) = J^{\gamma}_{(2)}(b_d,U)$.

The relations i) and ii) determine the constants b_d and $A_{d,1}$ and the details of the resulting algebra is identical with those yielding the equations (3.14) and (3.20) of Karatzas' paper [24] (see also section 5.2 above). Accordingly, we obtain:

$$\begin{cases} \nu g(b_d) = \gamma \\ A_{d,1} \quad = \frac{p'_d(b_d)}{\gamma^+} e^{\gamma^+ b_d} \end{cases} \tag{5.14}$$

with:

$$\nu g(x) = \int_0^{\infty} h^U(x + \frac{z}{\tilde{\gamma}}) e^{-z} dz. \tag{5.15}$$

Note that $A_{d,1}$ is strictly positive which imply that the smooth fitting at the boundary b_d is optimal.

Finally, the relations iii) implies:

$$A_{u,2} = \frac{\gamma}{\beta} e^{-\omega b_u} + \int_0^{b_u} \left[\frac{\beta h(s) + \lambda\gamma}{\beta U}\right] e^{-\omega s} ds \tag{5.16}$$

which, together with the relation iv), enable to calculate b_u. Note that there is no smooth fitting at the boundary b_u. This is due to the fact that the only possibility for the process $X(t)$ to reach b_u is to start at b_u (see [32] for details).

Lemma 5.4. *The function $\nu g(x)$ defined in equation (5.15) is strictly increasing.*

Proof: Integrating $\nu g(x)$ by parts we can write:

$$\nu g(x) = \int_0^{\infty} h(x + \frac{z}{\tilde{\gamma}}) e^{-z} dz\, (\rho - \gamma U) + \gamma U h(x). \tag{5.17}$$

Note that in equation (5.17) the function in the integral is on $h(x)$ and not $h^U(x)$. Hence, the derivative of $\nu g(x)$ reads as:

$$\nu g'(x) = \int_0^{\infty} h'(x + \frac{z}{\tilde{\gamma}}) e^{-z} dz\, (\rho - \gamma^- U) + \gamma U h'(x),$$

which is strictly positive. Indeed, $h(x)$ is strictly increasing and

$$\rho - \gamma^{-}U = \frac{2D\rho + \delta - \sqrt{\delta^2 - 4UD\rho\beta}}{2D} > 0. \tag{5.18}$$

The equation (5.18) follows directly from the definition of δ and ρ and the fact that $D < 0$.

□

Using Lemma 5.4 we conclude that the parameters b_d and b_u are the unique solutions of equations (5.14) and (5.16) respectively.

Using equation (5.7), note that the following properties consistently hold:

$$\lim_{x\to\infty} \nu g(x) = K \quad \text{and} \quad \lim_{x\to-\infty} \nu g(x) = k. \tag{5.19}$$

Theorem 5.5. *Assume that the instant reward function $h(x)$ is strictly increasing. Then for the initial conditions $X(0) = x$ and $I(0) = D$ respectively $I(0) = U$, the optimal stopping times $\tau^*(x, D, \gamma)$ respectively $\tau^*(x, U, \gamma)$ read as:*

a) *When $k < \gamma < K$:*

$$\tau^*(x, D, \gamma) = \inf\{t \geq 0; X(t) \leq b_d \text{ and } I(t) = D\} \tag{5.20}$$

and

$$\tau^*(x, U, \gamma) = \begin{cases} 0 & \text{if } x < b_u \\ \inf\{t \geq 0; X(t) \leq b_d \text{ and } I(t) = D\} & \text{if } x \geq b_u \end{cases} \tag{5.21}$$

where the thresholds b_d and b_u are implicitly given in equations (5.14) and (5.16). As the process $X(t)$ is increasing in the U phase and the reward function $h(x)$ is strictly increasing, remark that if $x \geq b_u$ we never stop while $I(t) = U$.

b) *When $\gamma > K$:*

In this case we have: $b_d = b_u = \infty$ hence $\tau^*(x, D, \gamma) = \tau^*(x, U, \gamma) = 0$. Therefore, the associated reward functions read as

$$J^{\gamma}(x, U) \equiv J^{\gamma}(x, D) \equiv J^{\gamma}_{(3)}(x, U) \equiv J^{\gamma}_{(3)}(x, D) = \frac{\gamma}{\beta}.$$

c) *When $\gamma < k$:*

In this cases we have: $b_d = b_u = -\infty$ and hence $\tau^*(x, D, \gamma) = \tau^*(x, U, \gamma) = \infty$. This implies that the associated reward functions read as:

$$J^{\gamma}(x, D) \equiv J^{\gamma}_{(1)}(x, D) = p_d(x) > k$$

and

$$J^{\gamma}(x, U) \equiv J^{\gamma}_{(1)}(x, U) = p_u(x) > k.$$

□

Remarks:

- When $0 \leq k < \gamma < K$, equations (5.20) and (5.21) implies that at the stopping time $\tau^*(x, D, \gamma)$ or $\tau^*(x, U, \gamma)$, the process $I(t)$ will always be in the D phase, namely:

$$I(\tau^*(x, D, \gamma)) = D \quad \text{or} \quad I(\tau^*(x, U, \gamma)) = D.$$

- By definition, if the initial condition is $(X(0) = x_0, I(0) = D)$ and we impose that $\gamma = \nu g(x_0)$, then immediate stopping is optimal (i.e. $\nu g(x)$ can be viewed as the Gittins index for the "down" state of the noise, see below for more details).

5.3.2 ⊛ *N*-Armed MABP

Now that the optimal stopping problem is solved let us derive the optimal solution for an N-armed MABP with piecewise deterministic dynamics:

$$\frac{d}{dt}X_j(t) = \mathbf{1}_{[\chi(t)=j]} I_j(t), \quad j = 1, 2, ..., N; \quad X_j(0) = x_j, \; I_j(0) = i_j \tag{5.22}$$

where the process $\chi(t) \in \{1, 2, \ldots, N\}$ indicates which project is engaged at time t and $\mathbf{1}_{[\chi(t)=j]}$ is the indicator function. As before, $I_j(t)$, $j \in \{1, 2, \ldots, N\}$ are independent alternating Markovian renewal processes similar to the one introduced in equation (5.5) with states U_j and D_j. Remember that when a project is not engaged its dynamics remains "frozen". Note that with the previous assumptions the processes $\zeta_k(t) := (X_k(t), I_k(t)) \in \mathbb{R} \times \{D_k, U_k\}$ for $k = 1, 2, \ldots N$ are independent and Markovian.

The MABP therefore consists in finding an optimal policy $\pi^* = \{\chi^*(t)\}$ which maximizes the expected reward:

$$J^{\pi^*}(\vec{X}, \vec{I}) = \max_{\pi \in \mathcal{U}} E\left\{ \int_0^\infty e^{-\beta t} h_{\chi^*(t)}\left(X_{\chi^*(t)}(t)\right) dt \;\middle|\; \vec{X}(t_0) = \vec{x}, \vec{I}(t_0) = \vec{i} \right\}, \tag{5.23}$$

where $\mathcal{U}$ is the set of admissible (i.e. non-anticipating) policy. In equation (5.23) we use the notation:

$$\vec{X}(t_0) = \left(x_1^0, x_2^0, ..., x_N^0\right) \quad \text{and} \quad \vec{I}(t_0) = \left(i_1^0, i_2^0, ..., i_N^0\right), \; i_k^0 \in \{U_k, D_k\},$$

The reward functions $h_j(x)$, for $j = 1, 2, ..., N$, are assumed to be strictly increasing, twice differentiable functions and, as before in equation (5.7), we impose that:

$$\lim_{x \to \infty} h_j(x) = K, \quad \lim_{x \to -\infty} h_j(x) = k, \;\; \forall\, j = 1, 2, \ldots N\,; \;\; 0 \le k < K.$$

In [34], (see also section 3.9 of [4]), it is shown that the optimal reward function equation (5.23) associated with the MABP for general Markov processes has a simple product form in terms of individual stopping problems without any smoothness properties of the optimal reward function neither for the global problem nor for the individual stopping problems. The processes $\zeta_j(t) := (X_j(t), I_j(t))$ being Markovian, the MABP defined by equations (5.22) and (5.23) belongs to the class discussed in [34]. For each individual process $X_j(t)$ given by equation (5.22), we then consider the associated stopping problem $\mathcal{P}_j$ from which we can define, for a given position x_j, two Gittins indices $\nu g_j^D(x_j)$ and $\nu g_j^U(x_j)$ depending on the respective states D and U of the of the noise as follows:

By definition, the Gittins indices are the smallest values of the terminal reward γ_j which make immediate stopping profitable when the state of the process is $\zeta_j = (x_j, D_j)$ respectively $\zeta_j = (x_j, U_j)$. In view of the results obtained above, the index $\nu g_j^D(x_j)$ therefore coincides with the expression equation (5.15). To identify the index $\nu g_j^U(x_j)$, we can conclude, from Figure 2, that it must be equal to the expected value of the stopping problem when the stopping reward is equal to $\nu g_j^D(x_j)$, i.e.

$$\nu g_j^U(x_j) = J^{\nu g_j^D(x_j)}(x_j, U_j)$$

Note in particular that both indices are monotonously increasing and that they obviously satisfy the relation equation (5.19).

The expressions $\nu g_k^D(x_k)$ and $\nu g_k^U(x_k)$ can be used as priority indices to characterize the optimal allocation policy π^* as follows: Define $\hat{\nu g}_j(x_j, I_j(t))$ as

$$\hat{\nu g}_j(x_j, I_j(t)) := \begin{cases} \nu g_j^D(x_j)) \text{ if } I_j(t) = D_j, \\ \\ \nu g_j^U(x_j) \quad \text{if } I_j(t) = U_j, \end{cases}$$

and the optimal strategy follows the rule: "*Play the project with the leading index* $\hat{\nu g}_j(x_j)$".

Observe that:

$$\nu g_j^D(x_j) \leq \nu g_j^U(x_j),$$

hence, for identical arms occupying a given identical position x_j, one should optimally engage the arm exhibiting the U_j state of the noise. This is obviously what is intuitively expected. Moreover, when $I_j(t) = U_j$ and project j is engaged, the process $X_j(t)$ is increasing and hence $\nu g_j^U(X_j(t))$ is increasing. Therefore it is optimal to continue engaging project j at least as long as the state of its noise is "up". In other words, all the disengaged projects X_k have $I_k(t) = D_k$.

5.3.3 ⊛ Illustration in the Manufacturing Context

In the manufacturing context, a simple illustration of PD processes can be given by considering a flexible failure-prone machine able to produce N different types of items. Due to its limited capacity, the machine can produce only a single item at a time. We assume the set-up costs and time to change from one production type to another to be negligible. As in [6] and in [17], we assume that the cumulated production $Y_j(t)$ of the items of class $j = 1, 2, \ldots, N$ can be described by a fluid equation of the form:

$$\frac{dY_j(t)}{dt} = I_j(t)\mathbf{1}_j(t), \quad Y_j(t) \geq 0, \quad \forall\, t, \tag{5.24}$$

where $\mathbf{1}_j(t) \in \{0, 1\}$ is the indicator which has the value 1 when the production of the item j is engaged and 0 otherwise, and $I_j(t) \in \{0, U_j\}$ is an alternating Markovian stochastic renewal process [42] with the assumption that the sojourn time in the "on", respectively "failed" phase is exponentially distributed with parameters λ_j, respectively μ_j.

Let us now introduce a set of constants D_j, $j = 1, 2, \ldots, N$ which are the target production rates for the item j. We can now write:

$$\frac{dX_j(t)}{dt} = [I_j(t) - D_j]\,\mathbf{1}_j(t), \quad j = 1, 2, \ldots, N. \tag{5.25}$$

We will interpret $X_j(t)$ as a performance measure of the production balance for item j at time t. Indeed from equation (5.25), we observe that when $X_j(t) \geq 0$, we fulfill or exceed the production target and conversely when $X_j(t) < 0$ we are below our objective. In order to be able to satisfy the demands on average we choose the D_j such that $D_j\lambda_j + U_j\mu_j > 0$. Note from equation (5.24), that when the production of item j is disengaged, both the time evolution of $X_j(t)$ and the state of the process $I_j(t)$ are "frozen".

When the production process j is engaged, we assume that an instantaneous performance gain $h_j(X_j(t))$ is achieved. The gain

$$h_j(X_j(t)) = h_j^0 + h_j^{extra}(X_j(t))$$

admits two contributions, namely a systematic contribution due to the intrinsic value of the item produced, say h_j^0, and an extra contribution $h_j^{\text{extra}}(X_j(t))$ which depends on the production balance $X_j(t)$. We shall assume that:

$$h_j^{\text{extra}}(X_j(t)) = 0 \quad \text{for} \quad X_j(t) \geq 0,$$

$$h_j^{\text{extra}}(X_j(t)) < 0 \quad \text{for} \quad X_j(t) < 0. \tag{5.26}$$

The contribution $h_j^{extra}(X_j(t))$ reflects that extra costs are incurred when the production is below the target rate. We assume that these costs increase monotonously with the distance to the target production and that

$$\lim_{x \to -\infty} h_j^{\text{extra}}(x) = -h_j^0. \tag{5.27}$$

From equations (5.26) and (5.27), we directly have:

$$0 \leq h_j(X_j(t)) \leq h_j^0 \tag{5.28}$$

and therefore $h_j(X_j(t))$ is monotonously increasing as required by equation (5.7).

Remark: Note that for a single class of items (i.e. when $N = 1$), equation (5.25) coincides (see [6] and [17]) with the time evolution of a surplus $X_j(t)$ with D_j being an external demand rate ($\mathbf{1}_j(t) = 1$ for all t in this case). However, when $N > 1$ the model defined by equation (5.25) differs fundamentally from the make-to-stock multiclass production context. Remember that for a make-to-stock multiclass production problem of the type described in [35] (see also part IV below), $X_j(t)$ represents the surplus of items of type j and equation (5.25) should be replaced by:

$$\frac{dX_j(t)}{dt} = I_j(t)\mathbf{1}_j(t) - D_j(t), \quad j = 1, 2, \ldots, N. \tag{5.29}$$

In view of equation (5.25) and equation (5.29) we observe that, for disengaged items, the demand continues to increase in the make-to-stock context (i.e. equation (5.29)), while $X_j(t)$ given by equation (5.25) remains frozen. Clearly, in equation (5.25), $X_j(t)$ is a performance measure which describes the operating characteristics of the machine when it delivers items of class j and this independently from the external demands.

6

Summary

It is established that for the classical MABP, the optimal scheduling rule is explicitly known (i.e. Priority Index Policy). We derived this policy in section 5.3 for a class of process describing the failure-prone machines (i.e. two-states random velocity models). As failure-prone machines are present in most manufacturing facilities, we would like to apply this result to actual production lines. Unfortunately, in most manufacturing scheduling problems for which the set-up costs and/or delays can be neglected, the classical MABP is still not directly applicable due to the "frozen" dynamics assumption. Clearly, when the scheduling problem depends on the external demands (as typically for the inventory levels), disengaged arms do evolve with time and the "frozen" dynamics hypotheses is violated. In general, the demands for the different types of items steadily increase independently of the fact that a particular production is engaged or not. The extended class of MABP for which the "frozen" assumption is relaxed is known as the Restless Bandit Problem (RBP). This class of dynamic scheduling problems will be studied in the next part.

Part II

Restless Bandit Problem (RBP)

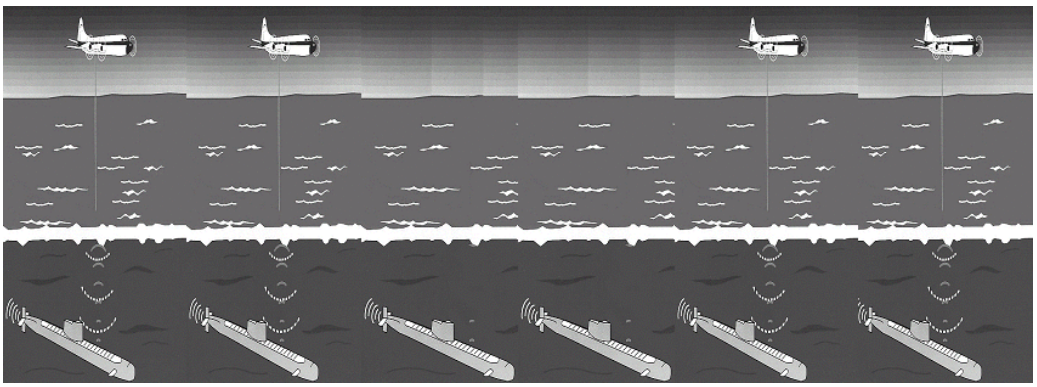

Fig. 6.1. Aircraft surveillance problem.

The Multi-Armed Bandit Problem (MABP) was suggested during the Second World War as an aircraft surveillance problem (see Figure 6.1):

"M aircrafts are trying to track the position of N enemy submarines with $M < N$. Accordingly the aircrafts must change task from time to time as all submarines are to be monitored. The problem is to allocate this surveillance in order to gain the maximal knowledge of the position of the submarine fleet."

In this problem, the projects (i.e. the knowledge of the position of the submarines) are restless in the most literal sense. Indeed, even untracked, the submarines continue to move, hence the "frozen" assumption characterizing the MABP is no more satisfied. The generalization of the MABP for which the "frozen" assumption is relaxed is known as the "Restless Bandit Problem" (RBP). We briefly present in chapter 7 the basic formulation of the continuous time, continuous state space version of the RBP, along the lines pioneered by Whittle [50]. We will neither define the discrete time version nor the continuous time and discrete state space version of the RBP as it follows naturally from the continuous time and continuous state space one. Nevertheless, the reader can find in section 8.1.3 an example of continuous time and discrete state space RBP explicitly solved.

The optimal solution for RBP is not yet known for general underlying dynamics and Priority index policies are known to yield sub-optimal results in general (see [49] and [47] for example). While the complete and analytical characterization of the optimal strategy for RBP remains a mathematical challenge, it is not clear that overcoming this difficulty will be of great benefit for actual applications. Therefore, in 1988 Whittle proposed a powerful heuristic for sub-optimally yet efficiently solving the RBP [50]. The Whittle heuristic is known under the name "Whittle relaxation" and is defined in section 8 below. Explicit expressions for the Whittle indices are derived under several underlying dynamics, including the diffusion dynamics in section 8.1.2 and the Markovian queue dynamics in section 8.1.3.

In part I the functions $h_j(x)$ was defined as being a reward function. The MABP can be indifferently defined with $h_j(x)$ being a cost function. When the $h_j(x)$ are cost functions, the aim of the Bandit problem is to minimize the global discounted cost rather than maximizing the global discounted reward. This cost problem is natural in the sense that in many applications we want to minimize the running cost (such as the production cost,

the storage cost, the setup cost...) We will assume in the sequel, that **the $h_j(x)$ are cost functions.** Therefore we will concentrate our efforts in order to **construct an efficient policy minimizing the global discounted cost of the RBP.**

Illustration of Restless Bandit Problems

The "frozen" assumption needed for the MABP is a very strong restriction. In many actual industrial problems this assumption precludes the utilisation of the MABP in order to model the problem. In this sense, the RBP is an important generalization of the MABP. Let us illustrate this through some simple examples of application:

The police control of drug markets problem [4]: In this problem, M police units are trying to control $N > M$ drug markets. The state of a project corresponds to the drug-dealing activity level of the corresponding drug market. When a police unit is focussed on an area, the drug-dealing activity is decreasing on this area (active phase). On the contrary, the drug-dealing activity is growing in the areas without police surveillance (passive phase). The goal is to move the police units in order to minimize the global drug-dealing activity.

Worker scheduling problem [50]: A number M of employees out of a pool of N have to be set to work at any time. The state of a project-worker represents his state of tiredness. Setting an employee to work (active phase) results in exhaustion of the corresponding worker, whereas letting it rest (passive phase) results in recuperation. The goal is to schedule the working time table of the pool in order to have workers as rested as possible.

The control of a make-to-stock production facility problem [35]: In this problem, a production facility can produce N different classes of items but it can produce only $M < N$ products at the same time. Each finished item is placed in its respective inventory, which services and exogenous demand. The level of the inventory represents the state of each project (item class). The goal is to minimize the costs due to the inventory and the costs due to the delay of delivery.

7

Restless Bandit Problem in Continuous Time and Continuous State Space – Definition

Consider a collection of N projects (i.e. N dynamical systems) evolving in the state space $\mathbb{R}$. The state of project j at time t is:

$$X_j(t) \in \mathbb{R}, \ \ j = 1, 2, ..., N.$$

We impose that at each instant $t \in \mathbb{R}^+$ **exactly $\mathbf{M < N}$ projects must be engaged** (i.e. must be in their active phase). If at time t, the project j is in state $X_j(t) = x_j$ and is engaged, then an **active running cost** $h_j^a(x_j)$ is incurred and the project evolves following an active transition probability. We assume in the following that the $X_j(t)$ are stationary Markovian stochastic processes and we use the notation

$$P_j\Big(X_j(t+dt) \ \Big| \ X_j(t), a\Big)$$

to describe the transition probability, where a indicates that the active action is selected. The running cost is discounted over time by a factor $e^{-\beta t}$. This means that the present value of one unit of tax equals $e^{-\beta t}$ when received t units of time in the future.

The other $N - M$ projects remain disengaged (i.e. remain in their passive phases). They generate **passive running costs** $h_k^p(x_k)$ and evolve according to stationary Markovian transition probabilities

$$P_k\Big(X_k(t+dt) \ \Big| \ X_k(t), p\Big),$$

where p indicates that the passive action is taken. The costs incurred in the passive phase are discounted by the same factor $e^{-\beta t}$.

Projects are to be selected for operation according to a *scheduling policy* $\pi \in \mathcal{U}_M$ where $\mathcal{U}_M$ is a subset of the admissible policies $\mathcal{U}$ (see Definition 3.1) having the property that each $\pi \in \mathcal{U}_M$ engages exactly M projects at each time t. We write

$$\vec{X}(t) = (X_1(t), \ldots, X_N(t))$$

for the state of the system at time t. The absence of switching penalties implies that the initial condition $\vec{u}_0 \in \{a, p\}^N$, describing the initial operation state (active or passive) of each project, is not necessary to characterize the evolution of $\vec{X}(t)$. Then the Restless Bandit Problem, for a given initial condition

$$\vec{X}(0) = \vec{x}_0 = (x_1^0, \ldots, x_N^0),$$

is to derive the optimal scheduling policy π^* which minimizes the total expected discounted cost $J^*(\vec{x}_0)$ over an infinite time horizon:

$$J^*(\vec{x}_0) = \inf_{\pi \in \mathcal{U}_M} E_{\vec{x}_0}^{\pi} \left[\int_0^{\infty} \sum_{k=1}^{N} h_j^{\Pi_j(s)}(X_j(s)) e^{-\beta s} ds \right]. \tag{7.1}$$

The superscript $\Pi_j(t) \in \{a, p\}$ denotes the operating state of the project j at time t when the policy π is adopted. The operator $E^{\pi}_{\vec{x}_0}$ denotes the expectation, under policy π, conditional to $\vec{x}_0$.

Let us define the action function:

$$I^{\pi}_j(t) = \begin{cases} 1 \text{ if the project } j \text{ is active at time t under policy } \pi, \\ 0 \text{ otherwise} \end{cases}$$

and

$$\bar{I}^{\pi}_j(t) = 1 - I^{\pi}_j(t).$$

In terms of $I^{\pi}_j(t)$ and $\bar{I}^{\pi}_j(t)$, we can rewrite the RBP as:

$$\text{RBP} \equiv \begin{cases} J^*(\vec{x}_0) = \\ \quad \inf_{\pi \in \mathcal{U}_M} E^{\pi}_{\vec{x}_0} \left[\underbrace{\int_0^{\infty} \sum_{j=1}^{N} h^a_j(X_j(t)) I^{\pi}_j(t) e^{-\beta t} dt}_{a)} + \underbrace{\int_0^{\infty} \sum_{j=1}^{N} h^p_j(X_j(t)) \bar{I}^{\pi}_j(t) e^{-\beta t} dt}_{b)} \right] \\ \\ \text{subject to the constraint} \\ \quad \sum_{j=1}^{N} I^{\pi}_j(t) = M, \quad \forall t \geq 0. \end{cases}$$

The term $a)$ describes the cost incurred by the active projects. Indeed, $I^{\pi}_j(t) = 0$ when project j is passive at time t. Similarly, the term $b)$ describes the cost incurred by the passive projects.

8

Heuristic Scheduling for the Restless Bandit Problem – The Whittle Relaxation

The complexity of the RBP has been shown in [7] to be PSPACE-hard. To solve this type of problems, one therefore relies in general on approximations. Whittle proposed in [50] an approximation scheme known as the *Whittle relaxation* problem (WR), which consists in relaxing the requirement that **exactly** M **projects** must be active at each time t, to the weaker requirement that M **projects must be active on average**. Accordingly, the WR reads as:

$$WR \equiv \begin{cases} J^W(\vec{x}_0) = \\ \quad \inf\limits_{\pi \in \mathcal{U}_W} E^{\pi}_{\vec{x}_0}\left[\int_0^\infty \sum\limits_{j=1}^{N} h_j^a(X_j(t)) I_j^\pi(t) e^{-\beta t} dt + \int_0^\infty \sum\limits_{j=1}^{N} h_j^p(X_j(t)) \bar{I}_j^\pi(t) e^{-\beta t} dt\right] \\ \text{subject to the constraint} \\ \quad E^{\pi}_{\vec{x}_0}\left[\int_0^\infty \sum\limits_{i=1}^{N} I_i^\pi(t) e^{-\beta t} dt\right] = \frac{M}{\beta}, \end{cases} \tag{8.1}$$

where $\mathcal{U}_W$ is a subset of the admissible policies $\mathcal{U}$ (see Definition 3.1) having the property that each $\pi \in \mathcal{U}_W$ engages M projects on average. Note that

$$\mathcal{U}_M \subset \mathcal{U}_W \subset \mathcal{U}.$$

From the definition of the action function $I_j^\pi(t)$, we have:

$$E^{\pi}_{\vec{x}_0}\left[\int_0^\infty \sum_{j=1}^{N} \bar{I}_j^\pi(t) e^{-\beta t} dt\right] = \frac{N-M}{\beta}.$$

Along the lines pioneered by Whittle, we use a Lagrangian multiplier to solve the problem (8.1). Accordingly, the Lagrange function $J^W(\vec{x}_0, \gamma)$ associated with equation (8.1) reads as:

$$J^W(\vec{x}_0, \gamma) = \sum_{j=1}^{N} J^j(x_j^0, \gamma) - (N-M)\frac{\gamma}{\beta}, \tag{8.2}$$

with

$$J^j(x_j^0, \gamma) = \inf_{\pi \in \mathcal{U}} E^{\pi}_{x_j^0}\Big[\int_0^\infty h_j^a(X_j(t)) I_j^\pi(t) e^{-\beta t} dt + \int_0^\infty (h_j^p(X_j(t)) + \gamma) \bar{I}_j^\pi(t) e^{-\beta t} dt\Big]. \tag{8.3}$$

Clearly the problem given in equation (8.1) is now decoupled into N single-project subproblems $J^j(x_j, \gamma)$ of the type given in equation (8.3). Following [50] and [36], we interpret the multiplier γ as playing the economic role of a constant *tax incurred when not producing*. Each single problem of the type arising in equation (8.3) is known as a γ-**penalty problem**.

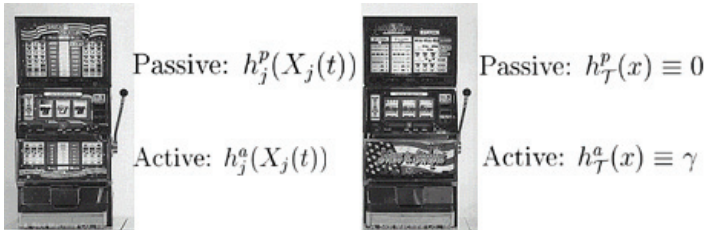

Fig. 8.1. Problem $\hat{\mathcal{P}}_j$.

Definition 8.1 (Problem $\hat{\mathcal{P}}_j$). *Given a N-armed RBP $\vec{X}(t)$, define N two-armed RBP (called problem $\hat{\mathcal{P}}_j$, $j = 1, \ldots, N$) as follows:*

The first arm of the two-armed RBP $\hat{\mathcal{P}}_j$, corresponds to the project j of $\vec{X}(t)$, with dynamics $X_j(t)$ and cost functions $h_j^\theta(X_j(t))$, $\theta \in \{a, p\}$. The second arm of $\mathcal{P}_j$ (called project $\mathcal{T}$) has the "frozen" dynamics:

$$X_{\mathcal{T}}(t) \equiv \xi \in \mathbb{R}, \ \forall \ t.$$

The project $\mathcal{T}$ gives a constant cost γ when activated and brings no cost when deactivated (i.e. $h_{\mathcal{T}}^a(x) \equiv \gamma$ and $h_{\mathcal{T}}^p(x) \equiv 0$, see Figure 8.1).

Remark: The γ-penalty problem $J^j(x_j^0, \gamma)$ is equivalent to the problem $\hat{\mathcal{P}}_j$. Indeed, at each instant of time, both the γ-penalty problem $J^j(x_j^0, \gamma)$ and the problem $\hat{\mathcal{P}}_j$ bring the same cost and have the same dynamics.

The γ-penalty problem belongs to the class of stationary Markovian decision problems [41]. It has been proven that for these problems, the optimal policy is stationary [41] (i.e. $\pi(X_j(t) = x_j)$ only depends on the state x_j of project j and not on the time t). Therefore, for each $x_j \in \mathbb{R}$ the optimal policy π^* either commands to activate project j or to let it passive (i.e. $\pi(x_j) \in \{a, p\}$). We can therefore make the following definition:

Definition 8.2 (Active States). *The set of active states $\mathcal{X}_j^a(\gamma) \subseteq \mathbb{R}$ contains all the states $x_j \in \mathbb{R}$ for which it is optimal to take the active action in the γ-penalty problem (8.3) (i.e. $\pi^*(x_j) = a$).*

Intuitively, the larger is the tax γ, the more state will belong to $\mathcal{X}_j^a(\gamma)$. Hence in terms of the set $\mathcal{X}_j^a(\gamma)$, the following structural property is essential:

Definition 8.3 (Indexability). *We say that problem (8.3) is indexable if the set $\mathcal{X}_j^a(\gamma)$ increases monotonically (i.e. $\gamma_1 < \gamma_2 \Rightarrow \mathcal{X}_j^a(\gamma_1) \subseteq \mathcal{X}_j^a(\gamma_2)$) from the empty set to the full state space as the tax γ increases from $-\infty$ to $+\infty$.*

Definition 8.4 (Index $\nu_j(x_j)$). *Assume that the problem $\hat{\mathcal{P}}_j$ (defined by equation (8.3)) is indexable. It follows that we can derive indices $\nu_j(x_j)$, for each $x_j \in \mathbb{R}$ such that the values $\nu_j(x_j)$ correspond to the smallest values of γ for which $x_j \in \mathcal{X}_j^a(\gamma)$.*

Remark: The indices $\nu_j(x_j)$ can be used to characterize the optimal solution of the γ-penalty problems for each fixed value of $\gamma \in \mathbb{R}$ as follows:

"Activate the project j when $\nu_j(X_j(t)) \leq \gamma$ at time t or let it be passive".

Hence, assuming that the project j is in the state x_j, the value $\nu_j(x_j)$ is the unique breakpoint of the tax γ which makes both the active and passive phase of project j indifferent for the γ-penalty problem. Equivalently, when $X_j(t) = x_j$, the index $\nu_j(x_j)$ is the smallest value of the tax γ that makes the immediate engagement of the "frozen" project $\mathcal{T}$ or the engagement of project j for the two-armed RBP $\hat{\mathcal{P}}_j$ indifferent.

Given an N-armed RBP $\vec{X}(t)$, we shall assume from now on, that the indexability of each problem $\hat{\mathcal{P}}_j$, $j = 1, \ldots, N$, holds. Under this assumption, we shall give a derivation of the indices $\nu_j(x_j)$, associated with the projects $j = 1, \ldots, N$ and we will construct an heuristic scheduling rule approaching the optimal policy of the RBP.

The Dynamic Programming (see [51] chapter 4, p.177), implies that the optimal cost function $J^j(x_j^0, \gamma)$ (defined in equation (8.3)) fulfills the property:

$$\min_{\theta \in \{a,p\}} \left[h_j^\theta(x_j^0) - \beta J^j(x_j^0, \gamma) + L(\theta) J^j(x_j^0, \gamma) \right] = 0, \tag{8.4}$$

where $L(\theta)$ is the infinitesimal generator of the controlled process $J^j(x_j^0, \gamma)$.

For notational ease, we define $J_\theta^j(x_j^0, \gamma)$ to be the solution of:

$$\left[h_j^\theta(x_j^0) - \beta J_\theta^j(x_j^0, \gamma) + L(\theta) J_\theta^j(x_j^0, \gamma) \right] = 0,$$

i.e. $J_\theta^j(x_j^0, \gamma)$ stands for the minimum discounted cost, when it is decided to take action $\theta \in \{a, p\}$ for the project j, at time 0.

As the indexability of the γ-penalized problem is assumed to holds, the index value $\nu_j(x_j^0)$ will be the minimal value of γ such that:

$$J_a^j(x_j^0, \gamma) = J_p^j(x_j^0, \gamma) \;\Leftrightarrow\; J_a^j(x_j^0, \nu_j(x_j^0)) = J_p^j(x_j^0, \nu_j(x_j^0)). \tag{8.5}$$

By deriving the index value for any initial condition $x_j \in \mathbb{R}$, we get the index function of project j:

$$\begin{aligned} \nu_j \;:\; & \mathbb{R} \to \mathbb{R} \\ & x_j \mapsto \nu_j(x_j) = \inf\{\gamma \in \mathbb{R} \mid \text{Eq.(8.5) holds}\}. \end{aligned} \tag{8.6}$$

This index will be call the Whittle Index in the following. Using the Whittle Index, we set the **generalized heuristic scheduling rule** for the multi-armed Restless Bandit problem, as:

Definition 8.5 (Whittle heuristic). *Assume that each project $j = 1, \ldots, N$, of an RBP is indexable, then the Whittle heuristic commands: "Engage at each time t the M projects exhibiting the M smallest index values $\nu_j(x_j)$ where $x_j = X_j(t)$, $j = 1, \ldots, N$".*

Remark: When γ is fixed, the optimal discounted cost $J^j(x_j, \gamma) = J_a^j(x_j, \gamma)$ for $x_j \in \mathcal{X}_j^a(\gamma)$ and $J^j(x_j, \gamma) = J_p^j(x_j, \gamma)$ for $x_j \notin \mathcal{X}_j^a(\gamma)$. In order to entirely define the optimal discounted cost $J^j(x_j, \gamma)$ in terms of $J_a^j(x_j, \gamma)$ and $J_p^j(x_j, \gamma)$, it remains to fit $J_a^j(x_j, \gamma)$ and $J_p^j(x_j, \gamma)$ on the active/passive boundary (i.e. the boundary of $\mathcal{X}_j^a(\gamma)$). In the following, we make use of the "smooth-fit" principle:

Definition 8.6 (Smooth-fit principle). *Suppose that x_j is on the active/passive boundary. Then the smooth-fit of $J_a^j(x_j,\gamma)$ and $J_p^j(x_j,\gamma)$ reads as:*

$$\begin{aligned} &J_a^j(x_j,\gamma) = J_p^j(x_j,\gamma), \\ &\tfrac{d}{dx}J_a^j(x,\gamma)\Big|_{x=x_j} = \tfrac{d}{dx}J_p^j(x,\gamma)\Big|_{x=x_j}, \\ &\tfrac{d^2}{dx^2}J_a^j(x,\gamma)\Big|_{x=x_j} = \tfrac{d^2}{dx^2}J_p^j(x,\gamma)\Big|_{x=x_j}. \end{aligned} \tag{8.7}$$

Remarks:

- The smooth-fit principle was first studied in detail in [5]. See also [45] and [28, Chap. 1 & 6], as well as [29, p.636] for a discussion of this principle and its history.

- When the evolution of the $X_j(t)$ are given by diffusion processes, the smooth-fit principle yields the optimum (see section 3.8 of [45]). This is not necessarily so for cases where non-diffusive processes occur. Nevertheless, the aim of the Whittle relaxation being to give a heuristic (possibly suboptimal) solving the RBP, we will **systematically use the smooth-fit principle**.

- For problems with discrete state space (for example $\mathbb{Z}$ i.e. $J_\theta^j(x_0,\gamma) : \mathbb{Z} \to \mathbb{R}$), we embed $\mathbb{Z}$ into $\mathbb{R}$ and consider $J_\theta^j(x_0,\gamma)$ as being a continuous function on $\mathbb{R}$ (i.e. $J_\theta^j(x_0,\gamma) : \mathbb{R} \to \mathbb{R}$). Using this continuous function, the smooth-fit principle will be extended to:

$$\begin{aligned} &J_a^j(x_j,\gamma) = J_p^j(x_j,\gamma), \\ &\Delta J_a^j(x,\gamma)\Big|_{x=x_j} = \Delta J_p^j(x,\gamma)\Big|_{x=x_j}, \\ &\Delta^2 J_a^j(x,\gamma)\Big|_{x=x_j} = \Delta^2 J_p^j(x,\gamma)\Big|_{x=x_j}. \end{aligned} \tag{8.8}$$

  where the derivatives Δ are define as:

$$\begin{aligned} &\Delta J_\theta^j(x_0,\gamma) := J_\theta^j(x_0,\gamma) - J_\theta^j(x_0-1,\gamma) \\ &\Delta^2 J_\theta^j(x_0,\gamma) := J_\theta^j(x_0+1,\gamma) - 2J_\theta^j(x_0,\gamma) + J_\theta^j(x_0-1,\gamma). \end{aligned}$$

  We will verify a posteriori (section 17.2 and appendix B) that this extended principle yields the optimal results for the Markov chain dynamics.

8.1 Explicitly Solved Examples

The explicit computations of the priority index for an arbitrary underlying stochastic process, is generally an elaborated exercise. Here we compute this index for several simple types of dynamics. Some of the expressions derived in this section will later be used in the production engineering context in part IV.

For general processes $X_j(t)$ and arbitrary cost functions $h_j^\theta(x)$, $\theta \in \{a,p\}$, the indexability of the γ-penalty problem (8.3) is not guaranteed. To progress we will assume in the following that the processes $X_j(t)$ and the cost functions $h_j^\theta(x)$ possess the required properties to ensure the indexability to hold. As only single-armed Bandits are considered in the following sections, we will omit the item index j.

8.1.1 ⊛ Simple Deterministic Project

As an introductive illustration, we derive the index for the discounted version of the deterministic problem presented in section 5 of [50].

Consider a continuous time project $X(t) \in \mathbb{R}$ satisfying

$$\frac{d}{dt}X(t) = f_{\theta(X(t))}(X(t)), \tag{8.9}$$

where $f_{\theta(X(t))}(X(t))$ equals $f_a(X(t))$ when the project is in state $X(t)$ and is in its active phase, and equals $f_p(X(t))$ when the project is in the state $X(t)$ and is in its passive phase. Let us also introduce two instantaneous cost functions $h_\theta(x)$, $\theta \in \{a, p\}$.

To derive the index $\nu(x)$ for this problem, we first solve the corresponding deterministic γ-penalty problem $J^j(x_j, \gamma)$:

Lemma 8.7. *Following the optimal policy π^* for the γ-penalty problem, equation (8.4) read as:*

$$\begin{cases} h_a(x) + f_a(x)\frac{d}{dx}J_a(x,\gamma) - \beta J_a(x,\gamma) = 0 \\ \qquad \text{if } \pi^*(X(0) = x) = a, \\ \hline h_p(x) + f_p(x)\frac{d}{dx}J_p(x,\gamma) - \beta J_p(x,\gamma) + \gamma = 0. \\ \qquad \text{if } \pi^*(X(0) = x) = p. \end{cases} \tag{8.10}$$

Proof: Assume that $\pi^*(X(0) = x) = a$, then the first order time expansion of equation (8.4) reads as:

$$J_a(x,\gamma) = \xi h_a(x) + (1 - \beta\xi)(J_a(x,\gamma) + \xi\frac{d}{dt}X(0)\frac{d}{dx}J_a(x,\gamma)),$$

After neglecting the terms of order $\mathcal{O}(\xi^2)$ in the above expansion, one gets the required result. A similar expression can be directly derived when $\pi^*(X(0) = x) = p$.

□

We solve equation (8.10) for $\frac{d}{dx}J_\theta(x,\gamma)$, $\theta \in \{a, p\}$ and obtain:

$$\frac{d}{dx}J_\theta(x,\gamma) = \begin{cases} \frac{\beta J_a(x,\gamma) - h_a(x)}{f_a(x)} & \text{in the active phase,} \\ \frac{\beta J_p(x,\gamma) - h_p(x) - \gamma}{f_p(x)} & \text{in the passive phase.} \end{cases}$$

Using the smooth-fit principle given in equation (8.7) and solving for γ we obtain:

$$\nu(x) = h_a(x) - h_p(x) + \frac{(f_p(x) - f_a(x))(f_p(x)h_a'(x) - f_a(x)h_p'(x))}{f_p(x)(f_a'(x) - \beta) - f_a(x)(f_p'(x) - \beta)}. \tag{8.11}$$

Observe that in the limit $\beta \to 0$, equation (8.11) consistently reduces to the result given in the Proposition 8 of [50] i.e.

$$\nu(x) = h_a(x) - h_p(x) + \frac{(f_p(x) - f_a(x))(f_p(x)h_a'(x) - f_a(x)h_p'(x))}{f_p(x)f_a'(x) - f_a(x)f_p'(x)}.$$

8.1.2 ⊛ RBP with Dynamics Driven by Diffusion Dynamics.

Consider now the situation where the project $X(t)$ is a diffusion process solving the stochastic differential equation:

$$dX(t) = \mu(t)dt + \sigma(t)dW(t), \tag{8.12}$$

with $dW(t)$ a White Gaussian noise process. The controlled drift $\mu(t)$ and variance $\sigma(t)$ terms read:

$$\mu(t) = \begin{cases} \mu_a > 0 & \text{if the active action is chosen,} \\ \tilde{\mu}_p < 0 & \text{if the passive action is chosen,} \end{cases}$$

respectively:

$$\sigma(t) = \begin{cases} \sigma_a > 0 & \text{if the active action is chosen,} \\ \sigma_p > 0 & \text{if the passive action is chosen,} \end{cases}$$

where $\mu_a, \mu_p, \sigma_a, \sigma_p$ are fixed constant. Define μ_p as the absolute value of $\tilde{\mu}_p$ (i.e. $\mu_p = |\tilde{\mu}_p|$). Using the Itô formula, equation (8.4) can be written in the form (see for example [24]):

$$\begin{cases} \frac{1}{2}\sigma_a^2 \frac{d^2}{dx^2} J_a(x,\gamma) + \mu_a \frac{d}{dx} J_a(x,\gamma) - \beta J_a(x,\gamma) + h_a(x) = 0, & \text{when } \pi_\gamma(x) = a, \\ \\ \frac{1}{2}\sigma_p^2 \frac{d^2}{dx^2} J_p(x,\gamma) - \mu_p \frac{d}{dx} J_p(x,\gamma) - \beta J_p(x,\gamma) + h_p(x) + \gamma = 0, & \text{when } \pi_\gamma(x) = p, \end{cases} \tag{8.13}$$

where $\pi_\gamma(x)$ gives the optimal action (active or passive) to take when in state x. The general solution of equation (8.13) in $\mathbb{R}$ is:

$$J_a(x,\gamma) = C_a^+ e^{-w_a^+ x} + C_a^- e^{w_a^- x} + S_a(x,\gamma)$$

and

$$J_p(x,\gamma) = C_p^+ e^{w_p^+ x} + C_p^- e^{-w_p^- x} + S_p(x,\gamma),$$

with the notations:

$$w_\theta^+ = \frac{\sqrt{\mu_\theta^2 + 2\beta\sigma_\theta^2} + \mu_\theta}{\sigma_\theta^2} > 0,$$

$$w_\theta^- = \frac{\sqrt{\mu_\theta^2 + 2\beta\sigma_\theta^2} - \mu_\theta}{\sigma_\theta^2} > 0$$

where C_θ^+ and C_θ^- are four integration constants and $S_\theta(x)$ are the particular solutions of equation (8.13) corresponding to engage [respectively disengage] the project X_k forever (see [24] or section 5.2 above). We obtain:

$$S_a(x,\gamma) = E_{(x,\gamma)} \int_0^\infty e^{-\beta t} h_a(x(t))\, dt =$$

$$= \frac{2}{\sigma_a^2 (w_a^+ + w_a^-)} \left[e^{-w_a^+ x} \int_{-\infty}^{x} h_a(y)\, e^{w_a^+ y} dy + e^{w_a^- x} \int_{x}^{\infty} h_a(y)\, e^{-w_a^- y} dy \right],$$

and

$$S_p(x,\gamma) = E_{(x,\gamma)} \int_0^\infty e^{-\beta t} (h_p(x(t)) + \gamma)\, dt =$$

$$= \frac{2}{\sigma_p^2 (w_p^+ + w_p^-)} \left[e^{-w_p^- x} \int_{-\infty}^{x} (h_p(y) + \gamma)\, e^{w_p^- y} dy + e^{w_p^+ x} \int_{x}^{\infty} (h_p(y) + \gamma)\, e^{-w_p^+ y} dy \right].$$

If $x_0 = +\infty$ the optimal discounted cost is attained by remaining passive forever. Similarly, when $x_0 = -\infty$ the optimal discounted cost is attained by remaining active forever. From the fact that $w_a^+ > 0$ and that $w_p^+ > 0$ this implies:

$$\lim_{x\to\infty} J_p(x,\gamma) = \lim_{x\to\infty} S_p(x,\gamma) \Rightarrow C_p^+ = 0$$

and

$$\lim_{x\to-\infty} J_a(x,\gamma) = \lim_{x\to-\infty} S_a(x,\gamma) \Rightarrow C_a^+ = 0.$$

Using the smooth-fit principle described in equation (8.7), we can write, after straightforward but lengthy algebra, the index $\nu(x)$ in the compact form:

$$\nu(x) = \frac{w_p^+}{w_a^- \sigma_a^2}\Big[\mathcal{I}_2^p \sigma_a^2 \frac{(w_p^+ - w_a^-)}{w_p^+} + \mathcal{I}_1^a \sigma_p^2 \frac{(w_p^- - w_a^+)}{w_a^+} + h_a(x)\sigma_p^2 - h_p(x)\sigma_a^2\Big]. \tag{8.14}$$

where:

$$\mathcal{I}_1^\theta = \int_0^\infty h_\theta(x - \frac{y}{w_\theta^+})\, e^{-y} dy$$

and

$$\mathcal{I}_2^\theta = \int_0^\infty h_\theta(x + \frac{y}{w_\theta^+})\, e^{-y} dy.$$

Remark: When $h_p(x) = 0$, $\mu_p = 0$ and in the limit $\sigma_p \to 0$, the RBP converges to the static Bandit problem where the passive project remains "frozen" and does not incur cost. In this limit, the index given by equation (8.14) converges to the value

$$\nu(x) = \int_0^\infty h_a(x - \frac{y}{w_a^+}) e^{-y}\, dy, \tag{8.15}$$

which directly corresponds to Karatzas' result [24], provided we reinterpret the reward problem in [24] as a cost problem (see also the equation (5.4) of section 5.2 above).

8.1.3 ⊛ RBP with Dynamics Driven by Continuous Time Markov Chains

Let us finally consider the case where the processes $X_k(t)$ is a birth and death process (i.e. a continuous-time and discrete state Markov chain for which $X_k(t) \in \mathbb{Z}$). Assume that the holding time between the transitions from the state x to $x+1$ is exponentially distributed with parameter μ_a when the active action is chosen and μ_p for the passive action. Conversely, for the transitions from the state x to $x-1$, the parameter is λ_a for the active action and λ_p for the passive action. We impose that $\mu_a > \lambda_a$ and that $\mu_p < \lambda_p$. In other words, the time average of the process $X(t)$ increases when active and decreases when passive. The associated running costs rates are $h_\theta(x)$, $\theta \in \{a, p\}$.

Lemma 8.8. *Under the above assumptions and following the optimal policy π^*, equation (8.4) takes the form:*

$$\begin{cases} \beta J_a(x,\gamma) = h_a(x) + \lambda_a J_a(x-1,\gamma) + \mu_a J_a(x+1,\gamma) - (\lambda_a + \mu_a) J_a(x,\gamma) \\ \qquad\qquad \text{if } \pi^*(X(0) = x) = a. \\ \hline \beta J_p(x,\gamma) = h_p(x) + \lambda_p J_p(x-1,\gamma) + \mu_p J_p(x+1,\gamma) - (\lambda_p + \mu_p) J_p(x,\gamma) + \gamma \\ \qquad\qquad \text{if } \pi^*(X(0) = x) = p. \end{cases} \tag{8.16}$$

Proof: Assume that $\pi^*(X(0) = x) = a$, then the first order time expansion of equation (8.4) reads as:

$$J_a(x,\gamma) = \xi h_a(x) + (1 - \beta\xi)\Big[\xi\lambda_a J_a(x-1,\gamma) + \xi\mu_a J_a(x+1,\gamma) + (1 - \xi\lambda_a - \xi\mu_a) J_a(x,\gamma)\Big].$$

Neglecting the terms of order $\mathcal{O}(\xi^2)$ in the above expansion yields the required result. A similar expression can also be derived when $\pi^*(X(0) = x) = p$.

□

The general solutions of equation (8.16) are:

$$J_a(x,\gamma) = C_{a^+}(w_a^+)^x + C_{a^-}(w_a^-)^x + S_a(x,\gamma),$$
$$J_p(x,\gamma) = C_{p^+}(w_p^+)^x + C_{p^-}(w_p^-)^x + S_p(x,\gamma),$$

with $C_{\theta+}$ and $C_{\theta-}$, $\theta \in \{a,p\}$, four integration constants,

$$w_\theta^+ = \frac{(\beta+\lambda_\theta+\mu_\theta)+\sqrt{(\beta+\lambda_\theta+\mu_\theta)^2-4\lambda_\theta\mu_\theta}}{2\mu_\theta}$$

$$w_\theta^- = \frac{(\beta+\lambda_\theta+\mu_\theta)-\sqrt{(\beta+\lambda_\theta+\mu_\theta)^2-4\lambda_\theta\mu_\theta}}{2\mu_\theta}$$

and $S_a(x,\gamma)$, $S_p(x,\gamma)$ being the particular solutions which correspond to remaining active [respectively passive] forever. We derive $S_a(x,\gamma)$ in the appendix C and obtain:

$$S_a(x,\gamma) = \frac{1}{\mu_a(w_a^+ - w_a^-)}\left\{h_a(x)+(w_a^-)^x \sum_{k=-\infty}^{x-1} h_a(k)(w_a^-)^{-k}+(w_a^+)^x \sum_{k=x+1}^{\infty} h_a(k)(w_a^+)^{-k}\right\}.$$

A computation along the same line yields:

$$S_p(x,\gamma) = \frac{1}{\mu_p(w_p^+-w_p^-)}\left\{(h_p(x)+\gamma)+(w_p^-)^x \sum_{k=-\infty}^{x-1}(h_p(k)+\gamma)(w_p^-)^{-k}+ (w_p^+)^x \sum_{k=x+1}^{\infty}(h_p(k)+\gamma)(w_p^+)^{-k}\right\}.$$

For consistency, it is required that the total cost incurred, when $x \to -\infty$ must equal the cost incurred when engaging the project forever. Moreover, when $x \to +\infty$ the global cost must equals the cost incurred when letting the project be idle forever. Using the fact that

$$0 \le w_\theta^- \le 1 \le w_\theta^+,$$

which is straightforward to establish, these asymptotic behaviours imply:

$$\lim_{x\to\infty} J_p(x,\gamma) = \lim_{x\to\infty} S_p(x,\gamma) \Rightarrow C_{p^+} = 0$$

and

$$\lim_{x\to-\infty} J_a(x,\gamma) = \lim_{x\to-\infty} S_a(x,\gamma) \Rightarrow C_{a^-} = 0.$$

Again, the index $\nu(x)$ is derived by fitting both functions $J_a(x,\gamma)$ and $J_p(x,\gamma)$ as described in equation (8.8). Explicit expressions for $\nu(x)$ will be given in section 17.2.1 below when a specific form of the cost function $h_\theta(x)$ is chosen.

9

Summary

The exact solution of stochastic scheduling problems involves the construction of a dynamic allocation policy in order to optimize a performance objective. This problem appears to be, in most relevant models, an unreachable goal. Therefore, the identification and study of restricted problems whose special structure yield a tractable solution remains of prime research interest. Beyond the intrinsic interest of writing a completely solvable model, its solution may provide building blocks for constructing well-grounded heuristic solutions for more complex models. In his article [50], Whittle studied what is arguably the most promising extension to the classical Multi-Armed Bandit Problem: the Restless Bandit Problem (RBP).

The rich natural modeling potential offered by the RBP in multiple disciplines spreading over robotics, aircraft surveillance, worker scheduling, make-to-stock queue or clinical trials, makes the development and analysis of a heuristic policy a problem of significant research importance. In his seminal paper on the subject, Whittle [50] presented a simple heuristic based on the Priority Index Policy. This class of heuristic is very appealing for actual problem as its implementation is easy and allows real time control. The remaining difficulty is to compute the indices. This can be done by using numerical computations or, in some simple cases, by deriving explicitly analytical results. The latter situation is exactly the one we applied in this chapter, where we derive explicit expressions for the Whittle priority indices when the underlying dynamics are diffusion or Continuous Time Markov Chains processes.

Part III

Multi-Armed Bandit Problem with Switching penalties

Introduction

So far in the MABP and RBP we have discussed neither costs nor time delays incurred when switching from one project to another one. This absence of switching penalties is barely encountered in the industrial context. In fact setup costs enter naturally in almost every production processes as cleaning operations or the need of additional workforce for setup. Think for example of a manufacturing process where a chocolate production is able to deliver plain chocolate P or chocolate with nuts N. For this problem, the setup time delay to go from P to N is almost negligible, on the contrary the setup to go from N to P will usually require a lengthy cleaning operation. Therefore, when the production N is engaged and demands for plain chocolate arrive, the Decision Maker (DM) will not necessarily switch immediately the production from N to P in order to avoid the switching penalties. Indeed, he may decide to continue to serve the demand for chocolate with nuts and wait until more demand for plain chocolate arrives.

This example shows the importance for the DM to know which is the production engaged at time t_{i-1} in order to take the optimal decision at time t_i. For a N-armed MABP, we will say that *"the DM is engaged on project j at time t_i"* if project j was engaged at time t_{i-1}. Therefore, if the DM is initially engaged on project j, this means that he can engage project j at time t_0 without incurring switching penalties. The information about the engaged project will be given by the indicator $I_j^\pi(t_i)$ defined by:

$$I_j^\pi(t_i) = \begin{cases} 1 & \text{if project } j \text{ is engaged at time } t_i \text{ under policy } \pi, \\ 0 & \text{otherwise.} \end{cases} \tag{9.1}$$

The objective of the present part is to study the difficulties arising when adding switching penalties into the MABP (defined in chapter 3). We will proceed as follows:

First, in chapter 10 we give a definition of the MABP with switching costs only. Then in chapter 11 we generalize the problem of chapter 10 and define the MABP including switching costs and/or switching times delays. In section 12.1 we will show that the presence of switching penalties in MABP, precludes the characterization of the optimal policy by using priority indices. To this end we will expose the counterexample constructed by Banks and Sundaram [3]. Nevertheless, numerical experiments such as those performed for instance in [22] and [38] show that in presence of switching costs, the optimal strategy exhibits a highly complex structure. This often implies complex implementations, a drawback that will drive most practitioners to prefer efficient (though sub-optimal) rules which are more easy to use. In particular, strategies based on generalized priority indices potentially remain –due to their simplicity– very appealing. We therefore propose in section 12.2 a possible generalization of the Priority Index policy based on two indices for each project.

How far from optimality can we expect to be when using generalized index policy in MABP with switching penalties? We will approach the answer of this question by studying a class of models involving MABP for which it is possible to exactly determine the optimal strategy by direct computation. The model we consider belongs to the class of deteriorating MABP (DMABP) for which the reward is monotonously decreasing. The DMABP plays a privileged role in the class of Bandit problems, as Kaspy and Mandelbaum in [27] proved that any MABP can be reduce to a Deteriorating one. Moreover it is a conjecture that the same construction affords to reduce any MABP with switching penalties into DMABP with switching penalties [26]. In chapter 13 we revue this construction and we characterize a class of MABP with switching costs for which it applies. Then we show in section 14.1 that the optimal policy for deterministic DMABP with switching penalties can be explicitly calculated and that when two arms are considered,

it exhibits an hysteretic shape. The hysteresis reflects the fact that not only the present state but also the history of the process is to be taken into account in order to decide which is the optimal scheduling. In chapter 15, we compare, for this two-armed process, the sub-optimal strategy resulting from the use of the GIH, with the optimal scheduling.

As we consider classical MABP in this part, **the functions $h_j(x)$ will be reward function** as in part I.

Fig. 9.1. MABP with Switching penalties.

10

Multi-Armed Bandit Problem with Switching costs

10.1 Definition in Discrete Time

We reconsider the class of models given in section 3.1 with $h_j(x),\ j \in \{1, 2, \ldots, N\}$ being reward functions and we now introduce the indicator:

$$\Delta^{\pi}(t_i) = \begin{cases} 1 & \text{if a switch occurs at time } t_i, \\ 0 & \text{otherwise,} \end{cases}$$

where π is the scheduling policy. As the knowledge of the engaged project is necessary to derive the optimal scheduling policy, the initial conditions are:

$$\vec{X}(t_0) = (X_1(t_0), \ldots, X_N(t_0)),$$

$$\vec{I}^{\pi}(t_0) = (I_1^{\pi}(t_0), \ldots, I_N^{\pi}(t_0)),$$

where $I_j^{\pi}(t_i)$ stands for the indicator function for the operation state (engaged or disengaged) of project j defined by equation 9.1.

Let us assume that the DM is initially engaged on project j, (i.e. $I_j(t_0) = 1$, and $I_{k\neq j} = 0$, $k \in \{1, \ldots, N\}$). Instead of equation 3.2, we now have to write for the optimal discounted reward $J^{\pi^*}(\vec{X}(t_0), \vec{I}(t_0))$:

$$J^{\pi^*}(\vec{X}(t_0), \vec{I}(t_0)) = \sup_{\pi \in \mathcal{U}} J^{\pi}(\vec{X}(t_0), \vec{I}(t_0))$$

where now

$$J^{\pi^*}(\vec{X}(t_0), \vec{I}(t_0)) =$$

$$\sup_{\pi \in \mathcal{U}} E_{\pi}\left\{ \sum_{i=0}^{\infty} \beta^{t_i} \underbrace{\left[\sum_{j=1}^{N} h_j(X_j(t_i)) I_j^{\pi}(t_i) \right]}_{a)} \underbrace{- C\Delta^{\pi}(t_i)}_{b)} \;\middle|\; \vec{X}(t_0),\ I_j^{\pi}(t_0) = 1 \right\}. \tag{10.1}$$

In equation 10.1, the term $a)$ represents the reward received by the engaged project (remember that $I_k^{\pi}(t_i) = 0$ for the disengaged project). The term $b)$ add the switching costs when a switching occurs at time t_i. Observe that the initial condition in equation 10.1 includes now the initial operating state $\vec{I}(t_0)$ of the MABP. This is mandatory as we have to take into account the fact that for arms being in identical dynamical states, the one currently engaged is more rewarding as a switching to the other one incurs a cost. As before

$$J^{\pi^*}(\vec{X}(t_0), \vec{I}(t_0)) = \max_k L_k J^{\pi^*}(\vec{X}(t_0), \vec{I}(t_0)) \tag{10.2}$$

with the one-step operator L_k introduced in equation 3.3 which takes now among one of the two alternatives:

The DM is engaged on project $j \neq k$ at time t_i and decide to switch to project k:

$$L_k J^\pi(\vec{X}(t_i), \vec{I}(t_i)) =$$
$$\Big[\underbrace{-C + h_k(X_k(t_i))}_{a)} + \underbrace{\beta E\Big[J^\pi(\vec{X}(t_{i+1}), \vec{I}(t_{i+1})) \mid \vec{X}(t_i), I_j^\pi(t_i) = 1\Big]}_{b)}\Big].$$

The term $a)$ is the switching cost incurred when switching from project j to project k plus the reward gained by the engagement of project k at time t_i. The term $b)$ describes the expected global reward from time t_{i+1} ahead discounted by the term β.

The DM is engaged on project k at time t_i and continue with project k:

$$L_k J^\pi(\vec{X}(t_i), \vec{I}(t_i)) =$$
$$\underbrace{h_k(X_k(t_i))}_{a)} + \underbrace{\beta E\Big[J^\pi(\vec{X}(t_{i+1}), \vec{I}(t_{i+1})) \mid \vec{X}(t_i), I_k^\pi(t_i) = 1)\Big]}_{b)}.$$

The term $a)$ describes the reward received by the engagement of project k at time t_i (here no switching occurs). The term $b)$ is as above.

10.2 Definition in Continuous Time

For continuous time, the definition of the MABP follows the one introduce in section 10.1 with the modifications:

- The $\{t_i,\ i = 0,\ 1, \ldots\}$, with $0 \leq t_1 < \cdots < t_i < t_{i+1} < \cdots,\ i = 1, 2, \ldots$, describe the sequence of ordered switching times.
- The discount factor in continuous time is $e^{-\beta t}$.
- Equation (10.1) is now rewritten as:

$$J^{\pi^*}(\vec{X}(0), \vec{I}^{\pi^*}(0)) =$$
$$\sup_{\pi \in \mathcal{U}} E_\pi \left\{ \int_0^\infty e^{-\beta t} \left(\sum_{j=1}^N h_j(X_j(t)) I_j^\pi(t) - C\delta^\pi(t - t_i) \right) dt \;\middle|\; \vec{X}(0), \vec{I}^\pi(0) \right\}, \qquad (10.3)$$

  with $\delta^\pi(t - t_i)$ being the Dirac mass distribution.

11

Multi-Armed Bandit Problem with Switching Costs and Switching Time Delays

Let us now consider the problem where **the switching penalties are costs and/or time delays**. This means that each time the DM decides to stop a project and engage another one, he has to wait a fixed time $D > 0$ before the new project becomes activated and pay a fixed tax $C > 0$. During this time delay, no project evolves and no reward is gained. Note that C and D depend neither on the project we leave nor on the project we engage.

11.1 Definition in Discrete Time

Remember that in discrete time, the t_i are the decision times. For simplicity, let us assume that $t_{i+1} - t_i = 1$ when no switching occurs at time t_i and $t_{i+1} - t_i = 1 + D$ when a switching occurs at time t_i. With the notations of section 10.1, the expected optimal reward $J^\pi(\vec{X}(t_0), \vec{I}(t_0))$ for the MABP with switching time delay reads as:

$$J^{\pi^*}(\vec{X}(t_0), \vec{I}(t_0)) =$$

$$\sup_{\pi \in \mathcal{U}} E_\pi \left\{ \sum_{i=0}^{\infty} \beta^{\,i + \sum_{j=0}^{t_i} \Delta^\pi(t_j) D} \left[\sum_{k=1}^{N} h_j(X_j(t_i)) I_j^\pi(t_i) - C \Delta^\pi(t_i) \right] \;\middle|\; \vec{X}(t_0),\ \vec{I}(t_0) \right\}. \quad (11.1)$$

As before,

$$J^{\pi^*}(\vec{X}(t_0), \vec{I}(t_0)) = \max_k L_k J^{\pi^*}(\vec{X}(t_0), \vec{I}(t_0)) \quad (11.2)$$

with the one-step operator L_k introduced in equation 3.3 which takes now among one of the two alternatives:

The DM is engaged on project $j \neq k$ at time t_i and decide to switch to project k:

$$L_k J^\pi(\vec{X}(t_i), \vec{I}(t_i)) =$$
$$\beta^D \Big[-C + h_k(X_k(t_i)) + \beta E \Big[J^\pi(\vec{X}(t_{i+1}), \vec{I}(t_{i+1})) \mid \vec{X}(t_i), I_j^\pi(t_i) = 1 \Big] \Big].$$

The DM is engaged on project k at time t_i and continue with project k:

$$L_k J^\pi(\vec{X}(t_i), \vec{I}(t_i)) =$$
$$h_k(X_k(t_i)) + \beta E \Big[J^\pi(\vec{X}(t_{i+1}), \vec{I}(t_{i+1})) \mid \vec{X}(t_i), I_k^\pi(t_i) = 1) \Big].$$

11.2 Definition in Continuous Time

For continuous time, the definition of the MABP with switching costs and switching time delays follows the one introduce in section 11.1 with the modifications:

- The $\{t_i,\ i = 0,\ 1, \ldots\}$, with $0 \leq t_1 < \cdots < t_i < t_{i+1} < \cdots,\ i = 1, 2, \ldots$, describe the sequence of ordered switching times.

- The discounted factor in continuous time is $e^{-\beta t}$.

- Equation (11.1) is now rewritten as:

$$J^{\pi^*}(\vec{X}(0), \vec{I}^{\pi^*}(0)) = \sup_{\pi \in \mathcal{U}} E_\pi \Bigg\{ \sum_{i=0}^{\infty} \left(e^{-\beta D}\right)^i \int_{t_i}^{t_{i+1}} e^{-\beta t} \left(\sum_{j=1}^{N} h_j(X_j(t)) I_j^\pi(t) - C\delta^\pi(t - t_i) \right) dt \ \Bigg| \ \vec{X}(0), \vec{I}^\pi(0) \Bigg\}, \tag{11.3}$$

  with $E_\pi\{ \cdot \mid \vec{X}(0), \vec{I}^\pi(0)\}$ being the conditional expectation with respect to the initial conditions and $\delta^\pi(t - t_i)$ is the Dirac mass distribution.

12

Heuristic Scheduling for the MABP with Switching penalties

The simple form of the scheduling policy given by priority indices strongly encourages us to look for a possible extension valid in presence of switching penalties. Following the procedure introduced in chapter 4, we can consider the decoupling of the original MABP into a series of N stopping problems $\mathcal{SP}_j$ and introduce a stopping cost. This method enables to construct N priority indices on which a scheduling policy can be based. Unfortunately such a naive extension leads generally to far from optimal policies. This should not come as a surprise as a simple decoupling of the MABP into N $\mathcal{SP}_j$ problems does not incorporate the information regarding the operating states (i.e. engaged or disengaged) of the projects. In presence of switching costs, this information is however essential as it is obvious that for two projects with neighbouring dynamical states, the currently engaged one is likely to be continued to avoid to pay a switching cost. This introductive idea suggests that hysteresis zone will enter into the scheduling diagram. The simplest manner to allow for the presence of these hysteresis is to introduce a set of two indices for each project $j \in \{1, 2, \ldots, N\}$, as follows:

- A "continuation index" $\nu c_j(X_j(t_i))$ for the case when the DM is engaged on the project j at time t_i.

- A "switching index" $\nu s_j(X_j(t_i))$, if project j was idle during the last period.

Given $\nu c_j(X_j(t_0))$ and $\nu s_j(X_j(t_0))$ for each project of an N-armed MABP, a generalization of the Gittins index policy for MABP with switching costs will be:

Definition 12.1 (Generalized Index Heuristic (GIH)). *Assume that the DM is initially engaged on project j, then the Generalized Index Heuristic (GIH) is: "Engage project j as long as $\nu c_j(X_j(t))$ is greater or equal than $\nu s_k(X_k(t))$, $\forall\, k \neq j$. If $\nu c_j(X_j(t))$ undergoes a switching index of another project, then switch to the project, different from project j, having the greatest switching index and engage it immediately".*

Given an N-armed MABP with switching penalties $\vec{X}(t)$, remember that $\mathcal{X}_j \in \mathbb{R}$ is the set of states space of project j. Define

$$\Theta = \mathcal{X}_1 \times \ldots \times \mathcal{X}_N \in \mathbb{R}^N,$$

then the GIH policy can be described by N^2 subsets:

$$S_{1\circlearrowleft}, S_{1\to 2}, S_{1\to 3}, \ldots, S_{1\to N} \subseteq \Theta$$

$$S_{2\circlearrowleft}, S_{2\to 1}, S_{2\to 3}, \ldots, S_{2\to N} \subseteq \Theta$$

$$\ldots$$

$$S_{N\circlearrowleft}, S_{N\to 1}, S_{N\to 2}, \ldots, S_{N\to N-1} \subseteq \Theta$$

as follows:

- The set $S_{j\circlearrowleft}$ contains the states $(x_1, \ldots, x_N) \in \Theta$ in which it is optimal to keep on engaging project j when the DM is already engaged on project j:

$$S_{j\circlearrowleft} = \Big\{(x_1, \ldots, x_N) \in \Theta \mid \nu c_j(x_j) \geq \nu s_k(x_k) \quad \forall k \in \{1, 2, \ldots, N\} \setminus \{j\}\Big\}.$$

- The set $S_{j\to k}$ contains the states $(x_1, \ldots, x_N) \in \Theta$ in which it is optimal to switch from project j to project k and to engage it immediately:

$$S_{j\to k} = \Big\{(x_1, \ldots, x_N) \in \Theta \mid \nu c_j(x_j) < \nu s_k(x_k) \quad \text{and } \nu s_k(x_k) = \max_{i=\{1,\ldots,N\}\setminus\{j\}} (\nu s_i(x_i))\Big\}.$$

Remarks:

- In figure 12.1 we draw the possible subsets describing the GIH for a particular two-armed MABP (in order to convince ourself that this situation may exist, read the chapter 14).

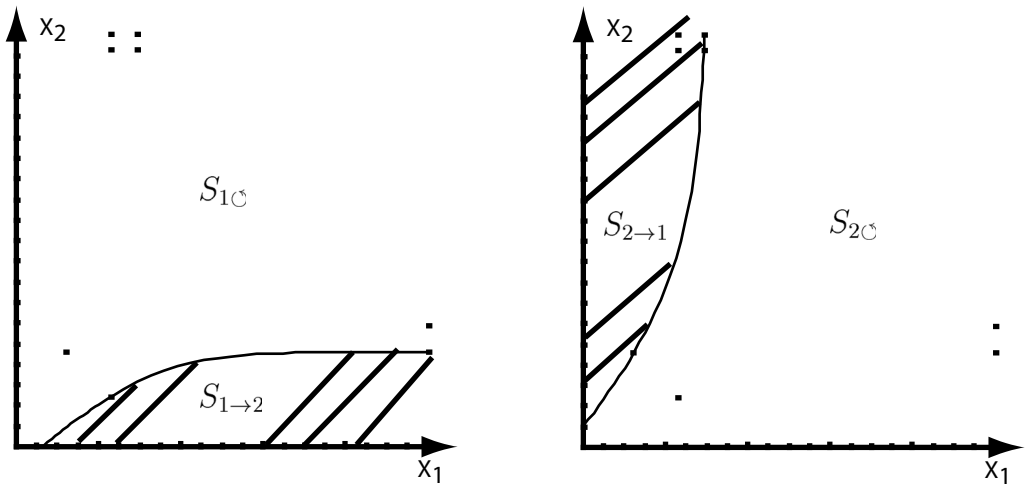

Fig. 12.1. Subsets describing the GIH - the two-armed case. The left graphic corresponds to the situation when the DM is initially engaged on project 1, and the right one corresponds to the situation when the DM is initially engaged on project 2.

- The boundary $\partial S_{j\to k}$ of the set $S_{j\to k}$ is generally different from $\partial S_{k\to j}$ so the GIH is a hysteretic policy.

12.1 The Counterexample of Banks and Sundaram

The following counterexample constructed by Banks, Rangarajan and Sundaram [3] shows that priority index policy based on $\nu c_j(x_j)$ and $\nu s_k(x_k)$ cannot possibly yield the optimal scheduling in the general case. Let us call $\mathcal{B}$ the class of Banks' MABP characterized by:

- The class $\mathcal{B}$ evolves in continuous state space with discrete time dynamics.
- The state space is $\mathcal{X} = [0,1] \in \mathbb{R}$.
- The initial state is $X(0) = x \in [0,1]$.
- The transition probabilities are $P(X(t+1) = 1 \mid X(t) = y) = y$, and $P(X(t+1) = 0 \mid X(t) = y) = (1-y)$.
- The reward function is $h(x) = x$.

This class of processes yields a nearly constant reward. Indeed, starting with initial condition x, it jumps either to position 1 with probability x and stays at 1 forever or it jumps to position 0 with probability $(1-x)$ and stays there forever. Following Banks, we will now prove that the optimal policy for MABP with switching costs $C > 0$, belonging to $\mathcal{B}$, cannot be a Priority Index Policy.

To show this, let us first assume, ad absurdum, that the GIH is optimal. Now, consider a two-armed MABP $\vec{X}_{\text{“Frozen”}}$ with switching costs $C > 0$. Suppose that the first project (project $\mathcal{T}_1$) of $\vec{X}_{\text{“Frozen”}}$ generates a systematic and constant reward $\gamma_1 \in \mathbb{R}$ and that its second project (project $\mathcal{T}_2$) generates a systematic and constant reward $\gamma_2 \in \mathbb{R}$. Therefore the dynamics of both projects of $\vec{X}_{\text{“Frozen”}}$ is “frozen” (i.e. $X_{\mathcal{T}_j}(t) \equiv \xi$, $t \in \mathbb{R}_+$), $j = 1, 2$. With these assumption we have:

Lemma 12.2. *The continuation and the switching index for a “frozen” project $\mathcal{T}$ with discrete time dynamics reads as:*

$$\nu c_{\mathcal{T}}(\xi) = \gamma \quad \text{and} \quad \nu s_{\mathcal{T}}(\xi) = \gamma - C(1-\beta). \tag{12.1}$$

Proof: Consider the MABP $\vec{X}_{\text{“Frozen”}}$. Then the optimal policy, if the DM is initially on project $\mathcal{T}_1$, is to continue forever on this project if and only if

$$\gamma_1 \geq \gamma_2 - C(1-\beta),$$

otherwise to switch to project $\mathcal{T}_2$ and stay on it forever. Hence, the GIH is optimal when it is based on a continuous and a switching index defined as in equation 12.1.

□

Consider now another two-armed MABP with switching costs $C > 0$. The dynamics of its first project (project $X_1(t)$) belong to $\mathcal{B}$ and the second project (called $\mathcal{T}$) is “frozen” and give a systematic and constant reward $\gamma \in \mathbb{R}$.

Assume that $2C(1-\beta) \leq x$ and that the DM is initially engaged on project X_1. Consider the two following alternative:

i) Engage project X_1 until it reaches the state 0, then switch to project $\mathcal{T}$ and engage it forever:

$$\underbrace{x}_{a)} + \underbrace{x \sum_{i=1}^{\infty} \beta^{t_i}}_{b)} + \underbrace{(1-x)\left(-\beta^{t_1} C + \sum_{i=1}^{\infty} \beta^{t_i} \gamma\right)}_{c)}. \tag{12.2}$$

The term a) describes the reward gained by engaging project 1 at time t_0. The term b) describes the reward gained by an infinite time engagement of project X_1 from time t_1 onward, under the assumption that it jumps from state x to state 1 (this occurs with probability x). The term c) describes the reward received by an infinite time engagement of project $\mathcal{T}$ from time t_1 onward with a switching at time t_1 (i.e. the

project X_1 jumps initially from state x to state 0, this occurs with probability $(1-x)$).

ii) Switch initially to the arm $\mathcal{T}$ and engage it forever:

$$\gamma \sum_{i=0}^{\infty} \beta^{t_i} - C. \tag{12.3}$$

Now, when

$$\gamma < \frac{x}{1-\beta(1-x)} + C(1-\beta), \tag{12.4}$$

we can easily check that the reward given by equation (12.2) is optimal. It is therefore greater than the reward given by equation (12.3). Moreover, when equation (12.4) is satisfied with equality, both rewards given by equations (12.2) and (12.3) are equivalent. Therefore, assuming that the GIH is optimal, we must have the following properties:

$$\begin{cases} \nu c_1(X_1(0)=x) > \nu s_{\mathcal{T}}(\xi) & \text{when } \gamma < \frac{x}{1-\beta(1-x)} + C(1-\beta) \\ \nu c_1(X_1(0)=x) = \nu s_{\mathcal{T}}(\xi) & \text{when } \gamma = \frac{x}{1-\beta(1-x)} + C(1-\beta) \\ \nu c_1(X_1(0)=x) < \nu s_{\mathcal{T}}(\xi) & \text{when } \gamma > \frac{x}{1-\beta(1-x)} + C(1-\beta) \end{cases}$$

Hence, for the GIH to be optimal, the continuation index of the class $\mathcal{B}$ have to be:

$$\nu c(X(0)=x) = \frac{x}{1-\beta(1-x)}. \tag{12.5}$$

Similar construction bring the switching index of the class $\mathcal{B}$, namely:

$$\nu s(X(0)=x) = \frac{x - C(1-\beta)(1+\beta(1-x))}{1-\beta(1-x)}. \tag{12.6}$$

Finally, consider another two-armed MABP for which both project belong to $\mathcal{B}$. Assume moreover that the following conditions are satisfied:

- The initial conditions are $X_1(0) = x \in [0,1]$ and $X_2(0) = y \in [0,1]$.
- The DM is initially engaged on the first project $X_1(t)$.
- $y > C(1-\beta) > x$.
- $\frac{x}{1-\beta(1-x)} > y - C(1-y)$.

With these assumptions, the optimal policy is: *"Engage the first project then switch to the second if and only if the state of the first project moves to 0 or continue with the first one forever"*. If the GIH is optimal, we must have

$$\nu c_1(x) > \nu s_2(y).$$

However, for $\beta = C = \frac{1}{2}$, $x = \frac{3}{17}$ and $y = \frac{27}{50}$ above assumption are satisfied and we have that $\nu c_1(x) = \frac{3}{10} = 0.3$ and $\nu s_2(y) = \frac{93}{308} \simeq 0.302$, hence a contradiction.

12.2 ⊛ Construction of the GIH

In order to be natural, the construction of the continuation and the switching index $\nu c_j(x)$ and $\nu s_j(x)$, must be very similar to the one of the Gitting index, which is as follows:

Given an N-armed MABP we derive $\nu g_j(x)$, from problems $\mathcal{P}_j$ $j \in \{1, 2, \ldots, N\}$. The $\mathcal{P}_j$ are stopping problems (see section 4.1) defined as follows: *"Given a terminal reward* $\frac{\gamma}{1-\beta}$*, find an optimal stopping time* $\tau^*(\gamma) \in \mathbb{R}_+$ *which maximizes the reward gained by engaging project* j *until time* $\tau^*(\gamma)$*, then stop and collect the terminal reward* $\beta^{\tau^*(\gamma)} \frac{\gamma}{1-\beta}$*."*

The simple formulation of the Gittins Index in term of a stopping problem is its major advantage. Moreover, it is worth to note that the Priority Index Policy based on the Gittins Index gives an optimal scheduling policy for the problems $\mathcal{P}_j$. We will therefore derive $\nu c_j(x)$ and $\nu s_j(x)$ directly from the stopping problem $\mathcal{P}_j$ in which we add switching penalties (i.e. a switching cost $C > 0$ and a switching time delay $D > 0$).

Remarks:

- The indices derived from the Whittle relaxation also follow from a stoping problem (problem $\hat{\mathcal{P}}_j$ see chapter 8).

- The Whittle heuristic (Definition 8.5) based on the Whittle Index (defined in equation (8.6)) gives an optimal scheduling policy for the problems $\hat{\mathcal{P}}_j$.

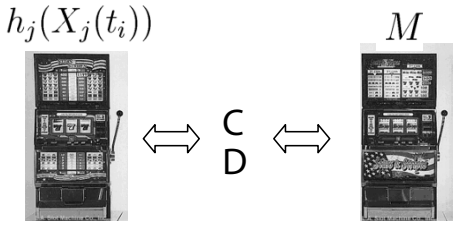

Fig. 12.2. Problem $\tilde{\mathcal{P}}_j$.

Definition 12.3 (Problem $\tilde{\mathcal{P}}_j$). *Given an N-armed MABP $\vec{X}(t_i)$ with switching penalties (a switching cost $C > 0$ and a switching time delay $D > 0$), define N two-armed MABP, called problems $\tilde{\mathcal{P}}_j$, as follows:*

The first arm of $\tilde{\mathcal{P}}_j$ corresponds to the project j of $\vec{X}(t_i)$, with dynamics $X_j(t_i)$ and yielding a reward $h_j(X_j(t_i))$. The second project of $\tilde{\mathcal{P}}_j$ (called project $\mathcal{T}$) has the "frozen" dynamics:

$$X_{\mathcal{T}}(t_i) \equiv \xi \in \mathcal{X}_{\mathcal{T}} \subset \mathbb{R}, \ \forall t_i$$

and produces a constant reward γ when engaged (i.e. $h_{\mathcal{T}}(\xi) \equiv \gamma$). Moreover, each time we switch from one project to another we incurred the switching penalties (see Figure 12.2).

Lemma 12.4. *The problem $\tilde{\mathcal{P}}_j$ is a stopping problem.*

Proof: Assume that the DM is initially engaged on the project j, then once the optimal policy for problem $\tilde{\mathcal{P}}_j$ commands to switch from project j to project $\mathcal{T}$, it is optimal to engage project $\mathcal{T}$ forever as the state of project j remains frozen. Then problem $\tilde{\mathcal{P}}_j$ is equivalent to the stopping problems (called as problem $\tilde{\mathcal{SP}}_j$) with stopping costs and stopping time delays defined as follows: *"Given a terminal reward* $\frac{\gamma}{1-\beta}$ *find an optimal stopping time* $\tau^*(\gamma) \in \mathbb{R}_+$ *which maximizes the reward gained by engaging project* j *until*

time $\tau^*(\gamma)$, then stop, wait D units of time, pay the switching tax C and collect the terminal reward

$$\beta^{\tau^*(\gamma)+D}\left(-C+\frac{\gamma}{1-\beta}\right)".$$

□

Definition 12.5 (Continuation index $\nu c_j(x_j)$). *Given an N-armed MABP $\vec{X}(t_i)$ with switching penalties (C,D), the continuation index $\nu c_j(x_j)$ of project j when in state x_j, is equal to the Gittins index (defined in equation (4.2)) i.e.*

$$\nu c_j(X_j(t_0)=x_j)=\sup_{\pi\in\mathcal{U}}\frac{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi-1}\beta^{t_i}h_j(X_j(t_i))\right\}}{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi-1}\beta^{t_i}\right\}}. \tag{12.7}$$

.

Remarks:

- We will see in section 12.2.2 that this definition directly follow from the problem $\tilde{\mathcal{P}}_j$.
- From definition 12.5, it directly follows that the Gittins index for the "frozen" project $\mathcal{T}$ reads as:

  $$\nu c_{\mathcal{T}}(\xi)=\gamma \tag{12.8}$$

  see lemma 4.8.

Definition 12.6 (Switching index $\nu s_j(x)$). *Given an N-armed MABP $\vec{X}(t_i)$ with switching penalties (C,D), the switching index $\nu s_j(x_j)$ of project j when in state x_j, is defined as the smallest value of γ that makes the immediate engagement of project $\mathcal{T}$ of problem $\tilde{\mathcal{P}}_j$ optimal (i.e. immediate stopping optimal) when the DM is initially engaged on project $\mathcal{T}$.*

Note that with the above definition, the optimal policy of problem $\tilde{\mathcal{P}}_j$, for each $\gamma < \nu s_j(x_j)$, commands to pay initially the switching penalties, then to engage project j until the optimal stopping time, finally to repay the switching penalties and to engage project $\mathcal{T}$ forever.

Lemma 12.7. *The switching index for the "frozen" project $\mathcal{T}$ with discrete time dynamics reads as:*

$$\nu s_{\mathcal{T}}(\xi)=\beta^D(\gamma-C(1-\beta)). \tag{12.9}$$

Proof: In order to derive the switching index for project $\mathcal{T}$ we construct (following the definition 12.6) a two-armed MABP with both projects having the "frozen" dynamics as defined for the project $\mathcal{T}$. Suppose that the first project (project $\mathcal{T}_1$) generates a constant reward γ_1 and that the second project (project $\mathcal{T}_2$) generates a constant reward γ_2. Then the optimal policy, if the DM is initially on project $\mathcal{T}_2$, is to continue forever on this project if and only if

$$\gamma_2\geq\beta^D(\gamma_1-C(1-\beta)),$$

otherwise to switch to project $\mathcal{T}_1$ and stay on it forever. Hence, the smallest value of γ_2 that makes immediate stopping optimal is

$$\gamma_2=\nu s_{\mathcal{T}_2}(\xi)=\beta^D(\gamma_1-C(1-\beta)). \tag{12.10}$$

□

Lemma 12.8. *The switching index for the "frozen" project $\mathcal{T}$ with continuous time dynamics reads as:*

$$\nu s_{\mathcal{T}}(\xi) = \beta^D(\gamma - \beta C). \tag{12.11}$$

Proof: Straightforward by following the lines given in the proof of Lemma 12.7.

□

Let us now derive the switching index for a general process $X_j(t)$ with a reward function $h_j(x)$.

12.2.1 ⊛ Construction of the Switching Index

Discrete Time Dynamics

Theorem 12.9 (The Switching Index). *Given an N-armed MABP $\vec{X}(t_i)$ with switching costs $C > 0$ and switching time delay $D > 0$, the switching index $\nu s_j(X_j(t_0))$ of project j reads as:*

$$\nu s_j(X_j(t_0)) = \sup_{\pi \in \mathcal{U}} \frac{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi - 1} \beta^{t_i} h_j(X_j(t_i)) - C(1 + \beta^{t_{\tau_\pi}+D})\right\}}{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi - 1} \beta^{t_i+D}\right\}}, \tag{12.12}$$

with t_{τ_π} the stopping time at which the policy π commands to engage the frozen project $\mathcal{T}$ of problem $\tilde{\mathcal{P}}_j$.

Proof: Given γ (not to large) the optimal reward $J_{\mathcal{T}}^{\gamma,C,D}(X_j(t_0), X_{\mathcal{T}} = \xi)$ for the problem $\tilde{\mathcal{P}}_j$ when the DM is initially engaged on project $\mathcal{T}$ can be written as:

$$\begin{aligned} J_{\mathcal{T}}^{\gamma,C,D}(X_j(t_0), X_{\mathcal{T}} = \xi) = \underbrace{\beta^D E\left[- C + \sum_{i=0}^{\tau_{\pi^*}-1} \beta^{t_i} h_j(X_j(t_i)) + \right.}_{a)} \\ \underbrace{\left. + \beta^D \left[-\beta^{t_{\tau_{\pi^*}}} C + \gamma \sum_{i=\tau_{\pi^*}}^{\infty} \beta^{t_i} \right] \right]}_{b)}, \end{aligned} \tag{12.13}$$

where τ_{π^*} is the time at which it is optimal to reengage $\mathcal{T}$. The term $a)$ describes the reward received by initially switching to project j and then engage it until time $t_{\tau_{\pi^*}}$. The term $b)$ describes the reward received from time $t_{\tau_{\pi^*}}$ onward when we switch from project j to project $\mathcal{T}$ at time $t_{\tau_{\pi^*}}$ and then we engage $\mathcal{T}$ forever.

When γ is the smallest value that makes immediate stopping optimal in the position $(X_j(t_0), \xi)$, the optimal reward is:

$$J_j^{\gamma,C,D}(X_j(t_0), \xi) = \frac{\gamma}{1-\beta} = E\left[\gamma \sum_{i=0}^{\infty} \beta^{t_i}\right]. \tag{12.14}$$

Using equations (12.14) and (12.13) we have:

$$E\left[\gamma \sum_{i=0}^{\infty} \beta^{t_i}\right] = \beta^D E\left[- C + \sum_{i=0}^{\tau_\pi - 1} \beta^{t_i} h_j(X_j(t_i)) + \beta^D \left[-\beta^{t_{\tau_\pi}} C + \gamma \sum_{i=\tau_\pi}^{\infty} \beta^{t_i} \right] \right],$$

which after simplification reads as:

$$\beta^D E\left[\gamma \sum_{i=0}^{\tau_\pi - 1} \beta^{t_i+D}\right] = \beta^D E\left[\sum_{i=0}^{\tau_\pi - 1} \beta^{t_i} h_j(X_j(t_i)) - C(1+\beta^{t_{\tau_\pi}+D})\right]. \tag{12.15}$$

Therefore, the smallest γ which make immediate engagement of project $\mathcal{T}$ optimal is solution of equation (12.15), namely:

$$\gamma = \sup_{\pi \in \mathcal{U}} \frac{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi - 1} \beta^{t_i} h_j(X_j(t_i)) - C(1+\beta^{t_{\tau_\pi}+D})\right\}}{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi - 1} \beta^{t_i+D}\right\}}. \tag{12.16}$$

By definition $\nu s_j(X_j(t_0)) = \gamma$ with γ given by equation (12.16), namely

$$\nu s_j(X_j(t_0)) = \sup_{\pi \in \mathcal{U}} \frac{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi - 1} \beta^{t_i} h_j(X_j(t_i)) - C(1+\beta^{t_{\tau_\pi}+D})\right\}}{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi - 1} \beta^{t_i+D}\right\}},$$

which finish the demonstration.

□

Continuous Time Dynamics

The continuous version of the switching index is similar to the discrete one:

Theorem 12.10 (The Switching Index). *The switching index $\nu s_j(X_j(t_0))$ reads as:*

$$\nu s_j(X_j(t_0)) = \sup_{\pi \in \mathcal{U}} e^{-\beta D}\left(\frac{E\left\{\int_0^{\tau_\pi} e^{-\beta t} h_j(X_j(t))\, dt - e^{-\beta D} C(1+e^{-\beta \tau_\pi})\right\}}{E\left\{\int_0^{\tau_\pi + 2D} e^{-\beta t} dt\right\}}\right) \tag{12.17}$$

with τ the stopping time at which the policy π engage the frozen project $\mathcal{T}$ of problem $\tilde{\mathcal{P}}_j$.

Proof: Proceed along the same lines as in the proof of Theorem 12.9, with equation (12.13) replaced by:

$$J_j^{\gamma,C,D}(X_j(t_0)) =$$

$$e^{-\beta D}\left(-C + E\left\{\int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\, dt - e^{-\beta(\tau^*+D)} C + \int_{\tau^*+D}^{\infty} \gamma e^{-\beta t} dt\right\}\right),$$

and equation (12.14) replaced by:

$$J_j^{\gamma,C,D}(X_j(t_0)) = \int_0^{\infty} \gamma e^{-\beta t} dt = \frac{\gamma}{\beta}.$$

□

12.2.2 ⊛ Construction of the Continuation Index

Let us now prove that the definition 12.5 for the continuation index is natural. To this end, we will prove that $\nu c_j(x)$ may be derived from the problem $\tilde{\mathcal{P}}_j$ and that the GIH based the continuation index (given in equation (12.7)) and the switching index (given in equation (12.9)) solve the problem $\tilde{\mathcal{P}}_j$ optimally.

Lemma 12.11. *The optimal policy for the problem $\tilde{\mathcal{P}}_j$, is the GIH policy based on the indices $\nu c_j(X_j(t))$ and $\nu s_{\mathcal{T}}(\xi)$, defined by equations (12.7) and (12.9.*

Proof of Lemma 12.11 for Discrete Time MABP

The optimal reward for the problem $\tilde{\mathcal{P}}_j$ when the DM is initially engaged on project j reads as:

$$J_j^{\gamma,C,D}(X_j(t_0),\xi) = E\left[\underbrace{\sum_{i=0}^{\tau_\pi-1}\beta^{t_i}h_j(X_j(t_i))}_{a)} + \underbrace{\beta^D\left[-\beta^{t_{\tau_\pi}}C+\gamma\sum_{i=\tau_\pi}^{\infty}\beta^{t_i}\right]}_{b)}\right], \tag{12.18}$$

where t_{τ_π} is the time at which it is optimal to engage project $\mathcal{T}$. The term $a)$ describes the reward received by initially engaging the project j until time t_{τ_π}. The term $b)$ describes the reward received from time t_{τ_π} onward when we switch from project j to project $\mathcal{T}$ at time t_{τ_π} and then we engage $\mathcal{T}$ forever.

For an initial condition $(X_j(t_0),\xi) \in \partial S_{j\circlearrowleft}$ and when the DM is initially engaged on project j, it is optimal to immediately switch to $\mathcal{T}$ and then to stay on it forever. This yields a reward:

$$J_j^{\gamma,C,D}(X_j(t_0)) = \beta^D(-C+\frac{\gamma}{1-\beta}) = \beta^D E\left[-C+\gamma\sum_{i=0}^{\infty}\beta^{t_i}\right]. \tag{12.19}$$

Using equation (12.19) with equation (12.18) we get:

$$\beta^D E\left[-C+\gamma\sum_{i=0}^{\infty}\beta^{t_i}\right] = E\left[\sum_{i=0}^{\tau_\pi-1}\beta^{t_i}h_j(X_j(t_i)) + \beta^D\left[-\beta^{t_{\tau_\pi}}C+\gamma\sum_{i=\tau_\pi}^{\infty}\beta^{t_i}\right]\right],$$

which yields after simplification:

$$\beta^D E\left[\gamma\sum_{i=0}^{\tau_\pi-1}\beta^{t_i}-C\right] = E\left[\sum_{i=0}^{\tau_\pi-1}\beta^{t_i}h_j(X_j(t_i)) - C\beta^{t_{\tau_\pi}+D}\right]. \tag{12.20}$$

On the other hand, if we want the GIH to be optimal for problem $\tilde{\mathcal{P}}_j$, we must have at the position $(X_j(t_0),\xi)$:

$$\nu c_j(X_j(t_0)) = \nu s_{\mathcal{T}}(\xi) = \beta^D(\gamma - C(1-\beta)), \tag{12.21}$$

with γ being the solution of equation (12.20), namely:

$$\gamma = \sup_{\pi\in\mathcal{U}}\beta^{-D}\frac{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi-1}\beta^{t_i}h_j(X_j(t_i)) + \beta^D C(1-\beta^{t_{\tau_\pi}})\right\}}{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi-1}\beta^{t_i}\right\}}. \tag{12.22}$$

Introducing equation (12.22) into equation (12.21) we obtain:

$$\nu c_j(X_j(t_0)) = \sup_{\pi\in\mathcal{U}}\frac{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi-1}\beta^{t_i}h_j(X_j(t_i))\right\}}{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi-1}\beta^{t_i}\right\}}, \tag{12.23}$$

which is the Gittins index. Therefore, for the GIH to be optimal with the switching index defined by equation (12.9), the continuation index has to be the Gittins Index.

□

Proof of Lemma 12.11 for Continuous Time MABP

The proof in continuous time follows naturally from the previous one:

The optimal reward $J_j^{\gamma,C,D}(X_j(t_0))$ for the problem $\tilde{\mathcal{P}}_j$ when the DM is initially engaged on arm j reads as:

$$J_j^{\gamma,C,D}(X_j(t_0)) = E\left\{ \int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\, dt - e^{-\beta(\tau^*+D)} C + \int_{\tau^*+D}^{\infty} \gamma e^{-\beta t} dt \right\}, \quad (12.24)$$

where τ^* is the time at which it is optimal to engage arm $\mathcal{T}$. For an initial condition $(X_j(t_0), \xi) \in \partial S_{j\circlearrowleft}$ and when the DM is initially engaged on project j, it is optimal to immediately switch to arm $\mathcal{T}$ and then to stay on it forever. This yields a reward:

$$J_j^{\gamma,C}(X_j(t_0)) = e^{-\beta D}\left(-C + \int_0^{\infty} \gamma e^{-\beta t} dt \right). \quad (12.25)$$

Using equation (12.25) into equation (12.24) we get:

$$\begin{aligned} & e^{-\beta D}\left(-C + \textstyle\int_0^{\infty} \gamma e^{-\beta t} dt \right) = \\ & E\left\{ \textstyle\int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\, dt - e^{-\beta(\tau^*+D)} C + \int_{\tau^*+D}^{\infty} \gamma e^{-\beta t} dt \right\}. \end{aligned} \quad (12.26)$$

On the other hand, if we want the GIH to be optimal for problem $\tilde{\mathcal{P}}_j$, we must have:

$$\nu c_j(X_j(t_0)) = \nu s_{\mathcal{T}}(\xi) = e^{-\beta D}(\gamma - C\beta), \quad (12.27)$$

with γ being the solution of equation (12.26), namely:

$$\gamma = e^{\beta D}\left(\frac{E\left\{ \int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\, dt + e^{-\beta D} C(1 - e^{-\beta \tau^*}) \right\}}{E\left\{ \int_0^{\tau^*} e^{-\beta t}\, dt \right\}} \right). \quad (12.28)$$

Introducing equation (12.28) into equation (12.27) we obtain:

$$\nu c_j(X_j(t_0)) = \frac{E\left\{ \int_0^{\tau^*} e^{-\beta t} h_j(X_j(t))\, dt \right\}}{E\left\{ \int_0^{\tau^*} e^{-\beta t}\, dt \right\}}, \quad (12.29)$$

which again is the Gittins index.

□

Remark: When $C \equiv 0$ and $D = 0$, we consistently have that $\nu s_j(X_j(t_0)) = \nu c_j(X_j(t_0)) = \nu g_j(X_j(t_0))$.

12.3 ⊛ Derivation of the GIH for the Banks' Class of MABP

To illustrate the above definition, let us derive the value of $\nu s_j(x)$ and $\nu c_j(x)$ for the class $\mathcal{B}$ of Bandit problems defined in section 12.1. For this class of MABP we can easily guess the optimal policy, therefore:

Proposition 12.12. *When $2C(1-\beta) \leq x$, the continuation index of a process which belongs to $\mathcal{B}$ is:*

$$\nu c(X(t_0)=x) = \frac{x}{1-\beta(1-x)} \tag{12.30}$$

and the switching index is:

$$\nu s(X(t_0)=x) = \beta^D \frac{x - C(1-\beta)(1+\beta^{D+1}(1-x))}{1-\beta^{2D+1}(1-x)} \tag{12.31}$$

Proof: Assume that the DM is initially engaged on project j and that $\gamma > 0$. Then the optimal policy of problem $\tilde{\mathcal{P}}_j$ for a process j in $\mathcal{B}$ is: *"Continue to engage project j until it reaches the state 0, then switch to project $\mathcal{T}$"*. Under this optimal policy, equation (12.20) reads as:

$$\underbrace{x}_{a)} + \underbrace{x\sum_{i=1}^{\infty}\beta^{t_i}}_{b)} + \underbrace{(1-x)\beta^D\left(-\beta^{t_1}C + \sum_{i=1}^{\infty}\beta^{t_i}\gamma\right)}_{c)} = \beta^D\left(\gamma\sum_{i=0}^{\infty}\beta^{t_i} - C\right). \tag{12.32}$$

The term a) describes the reward gained by engaging project j at time t_0. The term b) describes the reward gained by an infinite time engagement of project j from time t_1 onward, when with probability x it jumps from state x to state 1. The term c) describes the reward received when engaging project $\mathcal{T}$ forever from time t_1 onward, if at time t_1 we switch from project j to project $\mathcal{T}$ because the project j jumps from state x to state 0 (this occurs with probability $(1-x)$).

Solving equation (12.32) for γ we get:

$$\gamma = \beta^{-D}\frac{x}{1-\beta(1-x)} + C(1-\beta)$$

and according to Eq(12.10) the continuation index follows.

Similarly, with the assumption that the DM is initially engaged on project $\mathcal{T}$ and that γ is not too big, the optimal policy for a process j in $\mathcal{B}$ reads as: *"Switch initially to project j, engage it until it reaches the state 0, then switch to project $\mathcal{T}$ and engage it forever"*. Under this optimal policy, equation (12.15) reads as:

$$\beta^D\left(-C + x + x\sum_{i=1}^{\infty}\beta^{t_i} + (1-x)\beta^D\left(-\beta^{t_1}C + \sum_{i=1}^{\infty}\beta^{t_i}\gamma\right)\right) = \gamma\sum_{i=0}^{\infty}\beta^{t_i}.$$

Solving this equation for γ we get

$$\gamma = \beta^D\frac{x - C(1-\beta)(1+\beta^{D+1}(1-x))}{1-\beta^{2D+1}(1-x)}. \tag{12.33}$$

Finally, $\nu s_j(X_j(t_0)) = \gamma$ with γ given by equation 12.33.

□

Remark.:

- Observe that the equations (12.30) and (12.31) coincide with those derived in section 12.1 (i.e. equations (12.5) and (12.6)).

- The reason why the GIH is not optimal, comes from the fact that the GIH is "myopic regarding to the number of switchings" (i.e. we compute the indices by taking into account only one switching penalty ahead). Indeed, the indices $\nu s_j(x)$ and $\nu c_j(x)$ are derived by considering the problem $\tilde{\mathcal{P}}_j$. Hence, once it is optimal to switch from project j to project $\mathcal{T}$, it is never optimal to switch back to project j (as the state of j remains "frozen"). Hence we look only one switching ahead.

12.4 ⊛ Comparison Between the GIH and the Heuristic of Asawa and Teneketzis

In [2], the authors discuss the MABP with switching costs only (i.e. $D = 0$). They also propose a heuristic based on two indices, the Gittins index $\nu g_j(x)$ equation (4.3) and a switching index $\nu s'_j(x)$ defined as:

$$\nu s'_j(X_j(t_0)) = \sup_{\pi \in \mathcal{U}} \frac{E\left\{\sum_{i=0}^{\tau_\pi - 1} \beta^{t_i} h_j(X_j(t_i)) - C\right\}}{E\left\{\sum_{i=0}^{\tau_\pi - 1} \beta^{t_i}\right\}}. \tag{12.34}$$

With the formalism expose in chapter 4, their switching index is derived from the stopping problem ($\mathcal{SPA}_j$) defined as follows: *"Given a terminal reward $\frac{\gamma}{1-\beta}$ and an initial tax C to be paid prior to the engagement of project j, find an optimal stopping time $\tau^*(\gamma)$ which maximizes the total reward gained by engaging project j until time $\tau^*(\gamma)$, then stop and collect the reward $\beta^{\tau^*(\gamma)}\frac{\gamma}{1-\beta}$"*. Remember that our switching index $\nu s_j(x)$ (given in equation (12.12) is derived form the stopping problem $\tilde{\mathcal{SP}}_j$. The difference between these two stopping problems is that in $\mathcal{SPA}_j$ the DM does not pay a tax C at the stopping time $\tau^*(\gamma)$.

In [2], the authors established a series of Lemma that are satisfied by the indices $\nu s'_j(x)$ and $\nu g_j(x)$ for a two-armed MABP. Let us show that our GIH agrees with these Lemmas. The first result of [2] is *Lemma 2.7*:

Lemma 12.13 (Lemma 2.7 of [2]). *If at time t the DM is engaged on project j and $\nu g_j(X_j(t)) > \nu s'_j(X_j(t))$ then it is optimal to continue to engage project j.*

Proof: See [2]

□

In ou case we have:

Lemma 12.14. *The GIH based on the continuation and the switching indices equations (12.7) and (12.12) recommends to stay on project j for each state for which we have $\nu g_j(X_j(t)) > \nu s'_j(X_j(t))$.*

Proof: By definition, we have:

$$\nu s_j(X_j(t_0)) < \nu s'_j(X_j(t_0)),$$

so when $\nu g_j(X_j(t)) > \nu s'_j(X_j(t))$ the heuristic proposed by Asawa et al. [2] prescribes to keep on engaging project j and so does the GIH (as $\nu c_j(X_j(t)) > \nu s_j(X_j(t))$).

□

The second result is Lemma 2.8 of [2] which give a sufficient condition for the switching from one project to another:

Lemma 12.15 (Lemma 2.8 of [2]). *Consider a two-armed MABP with project j and project k. Assume that the DM is initially engaged on project j and that:*

$$\frac{E_{x_j}\left\{\sum_{i=0}^{\tau s_k^*-1}\beta^{t_i}h_j(X_j(t_i))-C(1+\beta^{t_{\tau_\pi}})\right\}}{E_{x_j}\left\{\sum_{i=0}^{\tau s_k^*-1}\beta^{t_i}\right\}}-\frac{E_{x_j}\left\{\sum_{i=0}^{\tau c_j^*-1}\beta^{t_i}h_j(X_j(t_i))\right\}}{E_{x_j}\left\{\sum_{i=0}^{\tau c_j^*-1}\beta^{t_i}\right\}}\geq\frac{2C\beta^{\tau c_j^*+\tau s_k^*}(1-\beta)}{(1-\beta^{\tau c_j^*})(1-\beta^{\tau s_k^*})}.$$

Here, τc_j^* and τs_k^* are the optimal stopping time for $\tilde{\mathcal{SP}}_j$, respectively $\tilde{\mathcal{SP}}_k$, with the DM is initially engaged on project j, respectively on project $\mathcal{T}$. It is then optimal to engage project k.

Proof: See [2]

□

With the GIH defined by equations (12.7) and (12.12), this lemma can be rewritten as:

Lemma 12.16 (Lemma 2.8 of [2]). *Consider a two-armed MABP with project j and project k. Assume that the DM is initially engaged on project j and that:*

$$\nu s_k(X_k(t_0))-\nu c_j(X_j(t_0))\geq\frac{2C\beta^{\tau c_j^*+\tau s_k^*}(1-\beta)}{(1-\beta^{\tau c_j^*})(1-\beta^{\tau s_k^*})}.$$

It is then optimal to engage project k.

□

Remarks:

- Computing the indices of Asawa et al. for the Banks' class of MABP (i.e. the class $\mathcal{B}$ defined in section 12.1), we find:

  $$\nu g(X(t_0)=x)=\frac{x}{1-\beta(1-x)}$$

  and

  $$\nu s'(X(t_0)=x)=\frac{x-c(1-\beta)}{1-\beta(1-x)}.$$

  As these expressions differs from the optimal one computed in section 12.1, they clearly yield a suboptimal scheduling for the problem $\tilde{\mathcal{P}}_j$.

- Does the indices of Asawa et al. yield a optimal reward for a two-armed MABP $\vec{X}_{\mathcal{B}}(t)$ with switching costs having both projects in the class $\mathcal{B}$? We ask this question because our indices gives only a suboptimal result in this case. In order to answer this question we will calculate the optimal policy of problem $\vec{X}_{\mathcal{B}}(t)$ and compare it with the scheduling obtained by the GIH based on the indices of Asawa et al. namely:

  $$\nu c_1(x)=\nu s_2'(y)\quad\text{and}\quad\nu c_2(y)=\nu s_1'(x)\tag{12.35}$$

  where $\nu s_i'(x)$ $i=1,2$ is define by equation (12.34).

12.4.1 ⊛ Optimal Policy for the Banks' Class of MABP with Switching Costs

As the project belonging to class $\mathcal{B}$ are Markovian and stationary, it is known (see [41], [42] or [43]) that the optimal policy for $\vec{X}_{\mathcal{B}}(t)$ belongs to the class of deterministic and stationary policies. Therefore, it can be defined by four subsets,

$$S_{1\circlearrowleft}, S_{2\circlearrowleft}, S_{1\to 2}, S_{2\to 1} \subseteq \mathcal{X}_1 \times \mathcal{X}_2$$

- The sets $S_{j\circlearrowleft},\ (j=1,2)$ contains the states $(x,y) \in \mathcal{X}_1 \times \mathcal{X}_2$ for which it is optimal to continue with project j when the DM is already engaged on it. Using equation (10.2), this set reads as:

$$S_{j\circlearrowleft} = \left\{ (x,y) \in \mathcal{X}_1 \times \mathcal{X}_2 \;\middle|\; L_j J_j^{\pi^*}(x,y) > L_k J_j^{\pi^*}(x,y),\ j \neq k \in \{1,2\} \right\}.$$

- The sets $S_{j\to_k},\ j \neq k \in \{1,2\}$ contains the states $(x,y) \in \mathcal{X}_1 \times \mathcal{X}_2$ for which it is optimal to switch from project j to project k when the DM is already engaged on project j. Using equation (10.2), this set reads as:

$$S_{j\to_k} = \left\{ (x,y) \in \mathcal{X}_1 \times \mathcal{X}_2 \;\middle|\; L_j J_j^{\pi^*}(x,y) \leq L_k J_j^{\pi^*}(x,y),\ j \neq k \in \{1,2\} \right\}. \quad (12.36)$$

On $\partial S_{1\to 2}$ we have that:

$$L_1 J_1^{\pi^*}(x,y) = L_2 J_1^{\pi^*}(x,y).$$

This implies that

$$\frac{x}{1-\beta} + (1-x)\left(\frac{y\beta}{1-\beta} - \beta C\right) =$$
$$-C + \frac{y}{1-\beta} + (1-y)\left(\frac{x\beta}{1-\beta} - \beta C\right).$$

We can explicitly solve these equations for y and:

$$\partial S_{1\to 2} = \left\{ (x, x + \frac{C}{1+\beta C}) \;\middle|\; x \in \mathcal{X}_1 \right\}.$$

Similar computations for $\partial S_{2\to 1}$ yield:

$$\partial S_{2\to 1} = \left\{ (y + \frac{C}{1+\beta C}, y) \;\middle|\; y \in \mathcal{X}_2 \right\}.$$

In Figure 12.3 we sketched the optimal switching cure as well as the switching curve obtained be equation (12.35). We clearly see that both switching curve are different. Therefore the indices of Asawa et al. are also suboptimal for problem $\vec{X}_{\mathcal{B}}(t)$.

In Figure 12.4 we sketched the optimal policy of problem $\vec{X}_{\mathcal{B}}(t)$ as well as the switching costs obtained by the GIH based on our indices i.e.

$$\nu c_1(x) = \nu s_2(y) \quad \text{and} \quad \nu c_2(y) = \nu s_1(x) \quad (12.37)$$

with the continuation and the switching index defined by equations (12.5) and (12.6) respectively. We see in this Figure that these boundaries are also different. Note nevertheless that the switching curve obtained be equation (12.37) are closer to the optimal one than the switching curve obtained by equation (12.35). The GIH based on our indices bring therefore better result than the heuristic of Asawa et al. in this case.

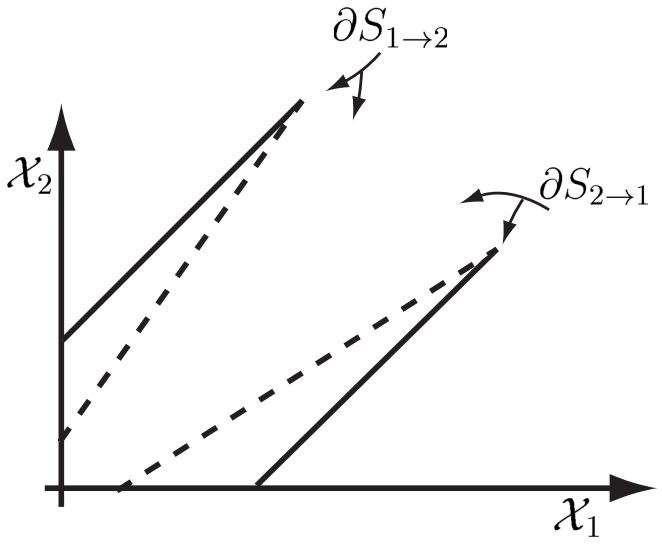

Fig. 12.3. The plain lines represent the optimal switching curve. The dashed lines represent the heuristic scheduling obtained following the Asawa's construction (see [2]).

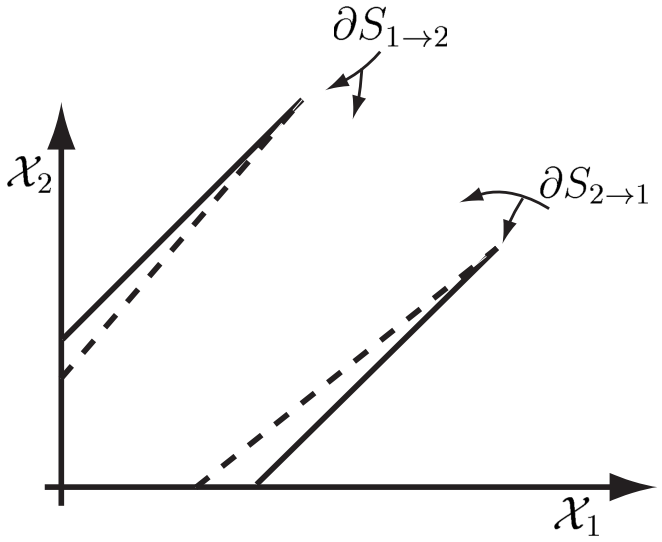

Fig. 12.4. The plain lines represent the optimal switching curve. The dashed lines represent the scheduling obtained by the indices equations (12.5) and (12.6).

13

Reduction of the MABP into the Deteriorating MABP.

In section 5.1, we considered the class of deteriorating MABP (DMABP), characterized by the fact that the processes $h_j(X_j(t))$ are non-increasing in time. The optimal scheduling policy for this class of Bandit reduces to engaging at each time the project having the largest instantaneous reward (i.e. a fully myopic policy and $\nu g_j(x) = h_j(x)$). Despite its simplicity the class of DMABP is important as any MABP can be reduced to a DMABP with an ad hoc construction [27]. This construction can be repeated for a class of MABP with switching costs [26] and we therefore get DMABP with switching costs. As the optimal policy for DMABP without switching cost is easy to derive, one expect that DMABP with switching costs offer the possibility to be solved. Let us start be deriving the construction of Kaspy which reduce the MABP with switching costs and discrete time dynamics into the DMAPB.

Let $\vec{X}(t)$ be an N-armed MABP with switching costs and reward function $h_j(x)$. Then the optimal global reward $J^{\pi^*}(\vec{X}(t_0), \vec{I}(t_0))$ for $\vec{X}(t)$ is given by equation (10.1),

$$J^{\pi^*}(\vec{X}(t_0), \vec{I}(t_0)) =$$

$$\sup_{\pi \in \mathcal{U}} E_\pi \left\{ \sum_{i=0}^{\infty} \beta^{t_i} \left[\sum_{j=1}^{N} h_j(X_j(t_i)) I_j^\pi(t_i) - C\Delta^\pi(t_i) \right] \;\middle|\; \vec{X}(t_0),\ I_j^\pi(t_0) = 1 \right\}.$$

The Kaspy's construction uses the Gittins $\nu g_j(x)$, $j \in \{1, 2, \ldots, N\}$ (given by equation (4.3)) and the following definition based on a realization of the random process $\nu g_j(X_j(t_i))$ (see figure 13.1).

- The lower envelope $\underline{\nu g}_j$ of νg_j is

$$\underline{\nu g}_j(X_j(t_i)) = \min_{0 \le t' < t_i} \nu g_j(X_j(t')).$$

- The strict decreasing ladder set is

$$\mathcal{M}_j = \{t_i \mid \nu g_j(X_j(t_i)) = \underline{\nu g}_j(X_j(t_i)) < \underline{\nu g}_j(X_j(t_{i-1}))\}$$

  i.e. we have that

$$\forall\ t_i \in \mathcal{M}_j,\quad \forall\ \tilde{t}_i < t_i,\ \nu g_j(X_j(t_i)) < \nu g_j(X_j(\tilde{t}_i)).$$

- Given $\gamma \in \mathbb{R}$, define

$$\tau_{t_i}^j(\gamma) = \inf\{\tilde{t}_i > t_i \mid \nu g_j(X_j(\tilde{t}_i)) < \gamma\}$$

  i.e. given a real value $\gamma \in \mathbb{R}$, we have that $\tau_{t_i}^j(\gamma)$ is the smallest decision time $\tilde{t}_i > t_i$ such that at time $\tilde{t}_i$ the index value of project j is smaller than γ.

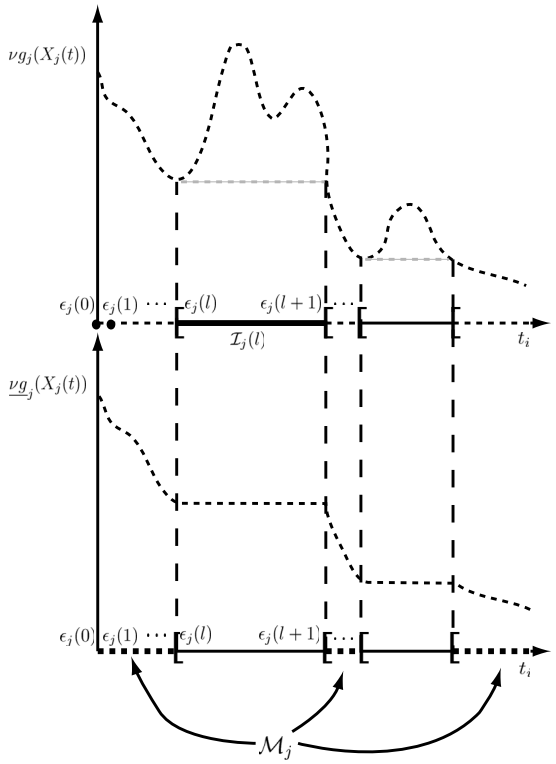

Fig. 13.1. One possible path realization of the Gittins Index.

– Let

$$\epsilon_j(0) = 0,\ \epsilon_j(l+1) = \inf\{t_i > \epsilon_j(l) \mid \nu g_j(X_j(t_i)) < \nu g_j(X_j(\epsilon_j(l)))\},$$

then the set $\mathcal{M}_j$, $j \in \{1, 2, \ldots, N\}$ may be defined as:

$$\mathcal{M}_j = \{\epsilon_j(l) \mid l = 1, 2, \ldots\}.$$

– Write the excursion intervals of project j as

$$\mathcal{I}_j(l) = \{t_i \in [\epsilon_j(l), \epsilon_j(l+1)[\mid I_j(t_i) = 1\}.$$

Remarks:

- From the definition of the excursion intervals we see that each $\epsilon_j(l)$ is the beginning of an excursion interval $\mathcal{I}_j(l)$ out of the strict decreasing ladder set.

- It is possible that the only point which belongs to an interval $\mathcal{I}_j(l)$ is $\epsilon_j(l)$. This is for example the case in the special realization drawn in figure 13.1 for each $\mathcal{I}_j(i)$ with $i < l$.

- If it exists a time $t_i \neq \epsilon_j(l)$ such that $t_i \in \mathcal{I}_j(l)$ then $\nu_j(X_j(t_i)) > \nu_j(X_j(\epsilon_j(l)))$.

- Remember that the indicator $I_j(t_i)$ represent the operation state of project j (i.e. $I_j(t_i) = 1$ if project j is engaged at time t_i and $I_j(t_i) = 0$ when it is disengaged). Hence, not all $t_i \in [\epsilon_j(l), \epsilon_j(l+1)[$ belong to the excursion intervals $\mathcal{I}_j(l)$. Indeed, the policy may command to switch from project j and engage another project at some times $t_k \in [\epsilon_j(l), \epsilon_j(l+1)]$, in this case $I_j(t_k) = 0$.

- Observe that

$$\nu g_j(X_j(\epsilon_j(l))) = \underline{\nu g}_j(X_j(\epsilon_j(l))) = \underline{\nu g}_j(X_j(t_i)); \ \forall\, t_i \in \mathcal{I}_j(l) \tag{13.1}$$

and

$$\tau^j_{\epsilon_j(l)}(\nu g_j(\epsilon_j(l))) = \epsilon_j(l+1), \ j \in \{1, 2, \ldots, N\}.$$

Indeed, $\forall\ t_i \in \mathcal{I}_j(l)$ we have that $\nu g_j(X_j(t_i)) \geq \nu g_j(\epsilon_j(l))$.

We also need the preliminary result:

Theorem 13.1.

$$\nu g_j(X_j(t_i)) = \frac{E_{x_j}\left\{ \sum_{i=0}^{\tau^j_{t_i}(\nu g_j(X_j(t_i)))-1} \beta^{t_i} h_j(X_j(t_i)) \right\}}{E_{x_j}\left\{ \sum_{i=0}^{\tau^j_{t_i}(\nu g_j(X_j(t_i)))-1} \beta^{t_i} \right\}},$$

Where the summation is performed until time $(\tau^j_{t_i}(\nu g_j(X_j(t_i))) - 1)$ *which corresponds to the smallest decision time* $\tilde{t}_i > t_i$ *for which we have*

$$\nu g_j(X_j(\tilde{t}_i)) < \nu g_j(X_j(t_i)).$$

Proof: The proof is given in [27].

□

Thanks to the previous definition, we now derive the reduction of the MABP into the DMABP. Writing equation (10.1) in which we omit, for ease of notation, the initial condition, we have:

$$J^{\pi^*}(\vec{X}(t_0), \vec{I}(t_0)) = \sup_{\pi \in \mathcal{U}} E_\pi \left\{ \sum_{i=0}^{\infty} \beta^{t_i} \left[\sum_{j=1}^{N} h_j(X_j(t_i)) I^\pi_j(t_i) - C\Delta^\pi(t_i) \right] \right\} =$$

$$\sup_{\pi \in \mathcal{U}} E_\pi \left\{ \sum_{j=1}^{N} \sum_{l=0}^{\infty} \sum_{t_i \in \mathcal{I}_j(l)} \beta^{t_i} h_j(X_j(t_i)) - \sum_{j=1}^{N} \sum_{l=0}^{\infty} \sum_{t_i \in \mathcal{I}_j(l)} \beta^{t_i} C\Delta^\pi(t_i) \right\}. \tag{13.2}$$

We shall now discuss the two sums separately.

$$E_\pi \left\{ \sum_{j=1}^{N} \sum_{l=0}^{\infty} \sum_{t_i \in \mathcal{I}_j(l)} \beta^{t_i} h_j(X_j(t_i)) \right\} = \sum_{j=1}^{N} \sum_{l=0}^{\infty} E_\pi \left\{ \sum_{t_i \in \mathcal{I}_j(l)} \beta^{t_i} h_j(X_j(t_i)) \right\} \tag{13.3}$$

By definition of the Gittins index given in equation 4.3, we have:

$$\frac{E_\pi \left\{ \sum_{t_i \in \mathcal{I}_j(l)} \beta^{t_i} h_j(X_j(t_i)) \right\}}{E_\pi \left\{ \sum_{t_i \in \mathcal{I}_j(l)} \beta^{t_i} \right\}} \leq \nu g_j(X_j(\epsilon_j(l))) \tag{13.4}$$

for all j and l (the left hand side do not archive the supremum). Inserting equation 13.4 in equation (13.3) and recalling equation (13.1), it follows that:

$$E_\pi \left\{ \sum_{j=1}^{N} \sum_{l=0}^{\infty} \sum_{t_i \in \mathcal{I}_j(l)} \beta^{t_i} h_j(X_j(t_i)) \right\} \leq E_\pi \left\{ \sum_{j=1}^{N} \sum_{l=0}^{\infty} \nu g_j(X_j(\epsilon_j(l))) \sum_{t_i \in \mathcal{I}_j(l)} \beta^{t_i} \right\} =$$

$$E_\pi\left\{\sum_{j=1}^{N}\sum_{l=0}^{\infty}\sum_{t_i\in\mathcal{I}_j(l)}\beta^{t_i}\underline{\nu g}_j(X_j(t_i))\right\}=E_\pi\left\{\sum_{i=0}^{\infty}\beta^{t_i}\sum_{j=1}^{N}\underline{\nu g}_j(X_j(t_i))I_j^\pi(t_i)\right\}. \qquad (13.5)$$

Lemma 13.2. *For each policy π which does not command to switch from project j to another project during an excursion interval $[\epsilon_j(l),\epsilon_j(l+1)[$, equation (13.5) is satisfied with equality.*

Proof: Each policy which does not command to switch during an excursion interval fulfills $\forall\ t_i\in[\epsilon_j(l),\epsilon_j(l+1)[,\ t_i\in\mathcal{I}_j(l)\big)$. Now, $\forall\ t_i\in\mathcal{I}_j(l)$, we have

$$\underline{\nu g}_j(X_j(t_i))=\nu g(X_j(\epsilon_j(l))).$$

Then

$$\sum_{t_i\in\mathcal{I}_j(l)}\beta^{t_i}\underline{\nu g}_j(X_j(t_i))=\nu g(X_j(\epsilon_j(l)))\sum_{t_i=\epsilon_j(l)}^{\epsilon_j(l)+\tau}\beta^{t_i}$$

with $\tau=\epsilon_j(l+1)-\epsilon_j(l)$. Finally, as the optimal policy does not commands to switch during an excursion interval we have, by definition of the Gittins index, that:

$$\nu g(X_j(\epsilon_j(l)))\sum_{t_i=\epsilon_j(l)}^{\epsilon_j(l)+\tau}\beta^{t_i}=\sum_{t_i\in\mathcal{I}_j(l)}\beta^{t_i}h_j(X_j(t_i)).$$

This last equation follows because τ is the optimal stopping time for an initial condition $X_j(\epsilon_j(l))$ ($\tau=\tau^j_{t_i}(\nu g_j(X_j(t_i)))-1$ and thus we can apply theorem 13.1).

□

Consider now the second sum in equation (13.2). This term contains all the switching costs payed during a realisation of $\vec{X}_j(t)$. By avoiding to count the switching costs incurred during an excursion interval and retaining only the switching costs incurred at the beginning of the excursion, we can write:

$$\sum_{j=1}^{N}\sum_{l=0}^{\infty}\sum_{t_i\in\mathcal{I}_j(l)}\beta^{t_i}C\Delta^\pi(t_i)\geq\sum_{j=1}^{N}\sum_{l=0}^{\infty}\beta^{\epsilon_j(l+1)}C\Delta^\pi(\epsilon_j(l+1)). \qquad (13.6)$$

Define

$$\tilde{\Delta}_j^\pi(t_i)=\begin{cases}1 & \text{if } \Delta^\pi(t_i)=1 \text{ and } t_i=\epsilon_j(l) \text{ for some } l=1,2,\ldots\\ 0 & \text{otherwise,}\end{cases}$$

Then equation (13.6) can be rewritten as:

$$\sum_{j=1}^{N}\sum_{l=0}^{\infty}\beta^{\epsilon_j(l+1)}C\Delta^\pi(\epsilon_j(l+1))=\sum_{j=1}^{N}\sum_{i=0}^{\infty}\beta^{t_i}C\tilde{\Delta}_j^\pi(t_i).$$

Here again, for each policy π which does not command to switch during an excursion interval $[\epsilon_j(l),\epsilon_j(l+1)[$, equation (13.6) is satisfied with equality (all the switching costs are counted).

Finally, equation (13.2) reads as:

$$J^{\pi^*}(\vec{X}(t_0),\vec{I}(t_0))\leq\sup_{\pi\in\mathcal{U}}E_\pi\left\{\sum_{i=0}^{\infty}\beta^{t_i}\left[\sum_{j=1}^{N}\underline{\nu g}_j(X_j(t_i))I_j^\pi(t_i)-C\tilde{\Delta}^\pi(t_i)\right]\right\}. \qquad (13.7)$$

The right hand side of equation (13.7) describes the reward gained by a MABP with dynamics $X_j(t_i)$ and reward functions $\underline{\nu g}_j(x)$, $j \in \{1, 2, \ldots, N\}$, which is therefore a DMABP.

Remarks:

- The equation (13.7) is satisfied with equality for any policy π which does not command to switch from project j to another project during an excursion interval $[\epsilon_j(l), \epsilon_j(l+1)[$.

- When $C = 0$, it is known that the optimal policy for a MABP is the Priority Index Policy based on the Gittins index. This policy does not switch during an excursion interval as it engaged at any time the project with the largest index value. Therefore, using the above reduction, any MABP with $C = 0$ is reduced to a DMABP for which the optimal policy is fully myopic.

- Each time the equation 13.7 is satisfied with equality, it is sufficient to know the optimal scheduling policy for the reduced DMABP in order to get the optimal policy for the original MABP. In other words, quoting Kaspy:

  [...] "*we may work with DMABP with reward* $\underline{\nu g}_j(X_j(t_i))$ *and restrict our attention to strategies that allow switching of projects only at the strict decreasing ladder point*" [...]

  Unfortunately, for MABP with switching costs, we do not know if all optimal policy preclude switching during an excursion interval. Nevertheless, for the following two classes of MABP with switching costs, the construction of Kaspy applied and equation 13.7 is satisfied with equality:

  - The class of MABP with switching costs for which the optimal policy does not switch during an excursion interval.

  - The class of MABP with switching costs which reduce to a DMABP with switching costs for which the optimal policy does not switch during an excursion interval.

14

Optimal Policy for a class of DMABP with Switching Costs.

14.1 ⊛ Deterministic DMABP with switching costs – The Two-Armed Case.

In this section, we focus on the optimal policy for the following class (class $\mathcal{Z}$) of two-armed DMABP with switching costs:

Definition 14.1 (The Class $\mathcal{Z}$). *The projects j belong to the class $\mathcal{Z}$ if:*

- *The dynamics of $X_j(t)$ is deterministic.*
- *The reward functions $h_j(x_j)$ is decreasing.*
- *Given $X_j(0)$, the instantaneous reward $h_j(X_j(t))$ fulfills:*

$$\lim_{t\to\infty} h_j(X_j(t)) = \Gamma_j \in \mathbb{R},\ \ j = 1, 2. \tag{14.1}$$

Theorem 14.2. *For a two-armed continuous time deterministic DMABP with switching costs, having both projects in the class $\mathcal{Z}$, the optimal policy is characterized by two non-decreasing switching curves $\mathcal{SO}_{1\to 2}$ and $\mathcal{SO}_{2\to 1}$. Moreover, for any initial condition, only a finite number of switchings occur under the optimal policy.*

Proof: The proof of theorem 14.2 lies on the three following propositions which are proven in the appendix E, F and G:

□

Proposition 14.3. *For any given initial condition, the optimal policy commands to switch only a finite number of times.*

Proposition 14.4. *The optimal policy is characterized by two switching curves $\mathcal{SO}_{1\to 2}$ and $\mathcal{SO}_{2\to 1}$ which can be respectively described by two functions, $\tilde{y}:\ x_1 \mapsto \tilde{y}(x_1)$ and $\tilde{x}:\ x_2 \mapsto \tilde{x}(x_2)$.*

Proposition 14.5. *The optimal switching curves $\mathcal{SO}_{1\to 2}$ and $\mathcal{SO}_{2\to 1}$ are non-decreasing.*

The above result are summarise in Figure 14.1 where we have plotted the optimal switching curves for our class $\mathcal{Z}$ of two-armed DMABP. In this figure, the increasing property of the switching curves can be seen explicitly.

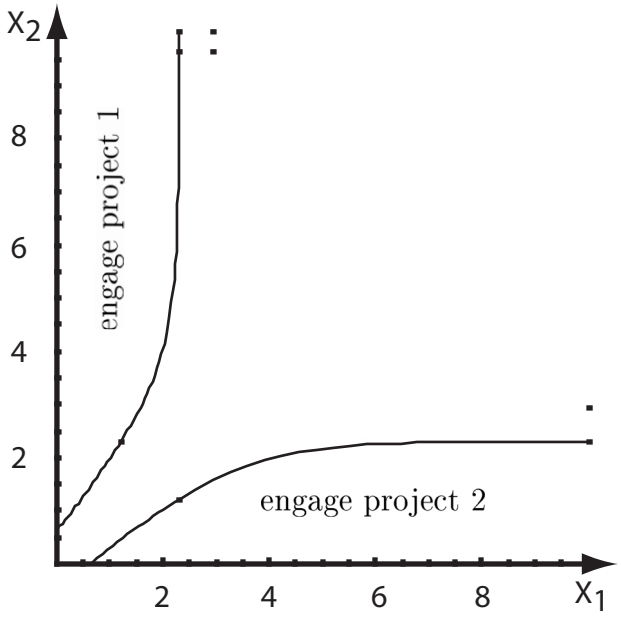

Fig. 14.1. Optimal policy for the class $\mathcal{Z}$ of two-armed DMABP.

14.2 ⊛ Explicit Derivation of the Switching Curves

As we consider Bandit problems with switching costs, the initial position of the Decision Maker (DM) is necessary to compute the global discounted reward. We will therefore include it, in the initial conditions and write them as:

$$(X_1(0), X_2(0), I)$$

where $I \in \{1, 2\}$ corresponds to the initial position of the DM. From the fact that the optimal switching curve $\mathcal{SO}_{1\to 2}$ is non-decreasing and the optimal policy involves only a finite number of switchings, follows the existence of a value A_1, such that for any initial condition $(X_1(0) \geq A_1, X_2(0), 2)$ the optimal policy commands to engage the project 2 forever. Similarly, it exists a value A_2, such that for any initial condition $(X_1(0), X_2(0) \geq A_2, 1)$, the optimal policy commands to engage the project 2 forever (i.e. the optimal switching curves exhibit the qualitative shape sketched in Fig.14.2a). We can compute these values as follows:

Starting with the initial condition $(\infty, A_2, 1)$, it is equivalent to either engage the project 1 forever, or to switch initially from project 1 to 2 and then engage it forever (i.e. the initial conditions $(\infty, A_2, 1)$ is on the switching curve). Accordingly, we can write:

$$\left[\int_0^\infty e^{-\beta t} h_1(X_1(t))\, dt \mid X_1(0) = \infty\right] = -C + \left[\int_0^\infty e^{-\beta t} h_2(X_2(t))\, dt \mid X_2(0) = A_2\right] \tag{14.2}$$

which determines A_2. In equation (14.2), we used the notation $[\,\cdot \mid X_j(t) = x_j]$ to indicate that the project j is in state x_j at time t. To simplify the exposition, we assume first that both projects have identical dynamics and reward characteristics (i.e. we consider symmetric DMABP). In this case, the strait line $x = y$ is an axis of symmetry in the Fig. 14.2a and hence $A_1 = A_2$.

The non-decreasing property of the switching curves allows us to determine them recursively. To see this, write

$$\begin{aligned} f \;:\; & \mathcal{X}_2 \to \mathbb{R} \\ & x_2 \mapsto f(x_2) \end{aligned}$$

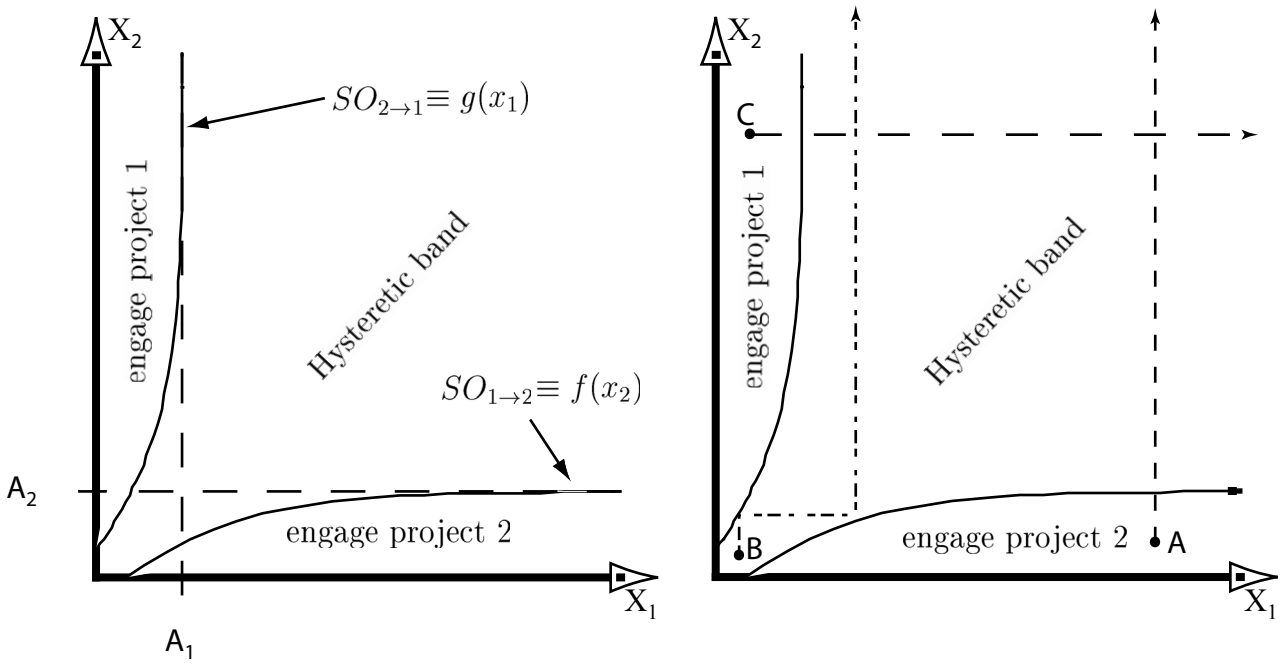

Fig. 14.2. a) Typical shape of the optimal policy. b) The dashed lines are the optimal trajectories for three different initial conditions A, B and C.

for the function which describes $\mathcal{SO}_{1\to2}$ and respectively

$$\begin{aligned} g \;:\; & \mathcal{X}_1 \to \mathbb{R} \\ & x_1 \mapsto g(x_1) \end{aligned}$$

for the function which describes $\mathcal{SO}_{2\to1}$. Define the sequences of points $(u_0, u_1, \ldots)$ and $(v_0, v_1, \ldots)$ as (see Fig. 14.3):

$$\begin{array}{ll} u_0 = A_1 & v_0 = A_2, \\ u_1 = g^{-1}(v_0) & v_1 = f^{-1}(u_0) \\ u_2 = g^{-1}(v_1) & v_2 = f^{-1}(u_1) \\ \quad\vdots & \quad\vdots \\ u_k = g^{-1}(v_{k-1}) & v_k = f^{-1}(u_{k-1}). \end{array}$$

Remark: For symmetric two-armed DMABP $g(x) = f^{-1}(x)$.

Iteration 1), computation of $\mathcal{SO}_{2\to1}$ in the interval $[u_1, A_1]$:

Assume that the DM is initially engaged on project 2, and that the initial positions are

$$u_1 \le X_1(0) = x_1 < A_1 \text{ and } X_2(0) = A_2$$

(as in Fig. 14.3). Following the optimal policy, the DM switches only once, when the state of the system reaches the position $(X_1(t) = x_1, X_2(t) = \bar{x}_2, 2)$ (i.e. $(x_1, \bar{x}_2)$ lies on $\mathcal{SO}_{2\to1}$, see Fig. 14.3). Therefore the optimal reward for the initial condition $(x_1, A_2, 2)$ satisfies:

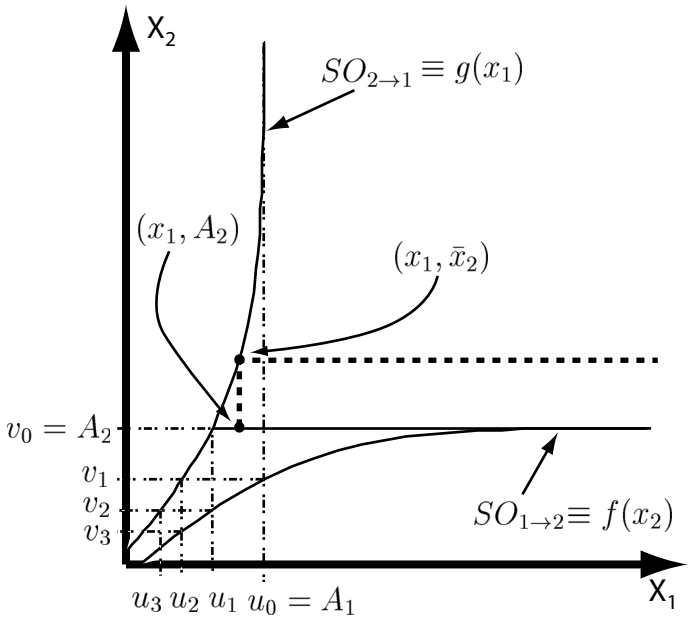

Fig. 14.3. Optimal policy for an initial condition $u_1 < x_1 \leq A_1$.

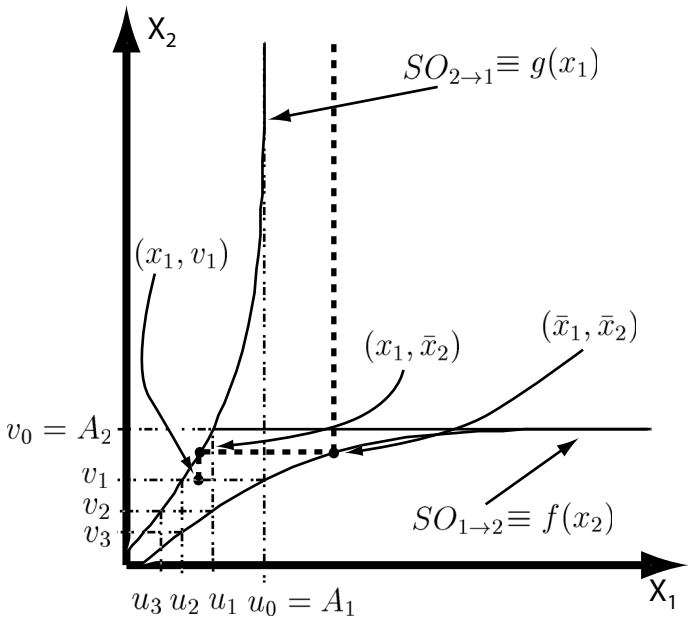

Fig. 14.4. Optimal policy for an initial condition $u_2 < x_1 \leq u_1$.

$$JO(x_1, A_2, 2; \bar{x}_2) = \underbrace{\left[\int_0^{\tau(\bar{x}_2)} e^{-\beta t} h_2(X_2(t))\, dt \mid X_2(0) = A_2\right]}_{a)} + \underbrace{e^{-\beta\tau(\bar{x}_2)} \left(-C + \left[\int_0^{\infty} e^{-\beta t} h_1(X_1(t))\, dt \mid X_1(0) = x_1\right] dt\right)}_{b)}, \tag{14.3}$$

where $\tau(\bar{x}_2)$ is the smallest time at which the process $X_2(\tau(\bar{x}_2)) = \bar{x}_2$ (i.e. X_2 is on $\mathcal{SO}_{2\to1}$). The term $a)$ describes the reward received by engaging the project 2 until time

$\tau(\bar{x}_2)$. The term b) describes the reward received when we switch from project 2 to project 1 at time $\tau(\bar{x}_2)$ and then we engage project 1 forever.

By optimality, the value of $\bar{x}_2$ must satisfies:

$$\frac{\partial}{\partial \bar{x}_2} JO(x_1, A_2, 2; \bar{x}_2) = 0.$$

For the symmetric DMABP, we directly get the switching curve $\mathcal{SO}_{1\to 2}$ on the interval $[A_1, \infty]$ by symmetry. Now we can compute the position of the switching curve $\mathcal{SO}_{2\to 1}$ on the interval $[u_2, u_1]$ as follows:

Iteration 2), computation of $\mathcal{SO}_{2\to 1}$ in the interval $[u_2, u_1]$:

Assume that project 2 is initially engaged and that the initial positions are

$$u_2 \le X_1(0) = x_1 < u_1 \text{ and } X_2(0) = v_1.$$

Following the optimal policy, the DM will switch exactly twice, first in the interval $[u_2, u_1]$, when the state of the system reaches the position $(X_1(t) = x_1, X_2(t) = \bar{x}_2, 2)$ and a second times in the interval $[A_1, \infty]$ when the state of the system reaches the position $(X_1(t) = \bar{x}_1, X_2(t) = \bar{x}_2, 1)$ (note that $\mathcal{SO}_{1\to 2}$ for $x \in [u_1, A_1]$ has been computed previously, see Fig.14.4). Therefore the optimal reward for $(x_1, v_1, 2)$ is:

$$\begin{aligned} JO(x_1, v_1, 2; \bar{x}_2) = & \underbrace{\left[\int_0^{\tau_1(\bar{x}_2)} e^{-\beta t} h_2(X_2(t))\, dt \;\middle|\; X_2(0) = v_1 \right]}_{a)} + \\ & \underbrace{e^{-\beta \tau_1(\bar{x}_2)} \Big(-C + \left[\int_0^{\tau_2(\bar{x}_1)} e^{-\beta t} h_1(X_1(t))\, dt \;\middle|\; X_1(\tau_1(\bar{x}_2)) = x_1 \right]}_{b)} + \\ & \underbrace{e^{-\beta(\tau_1(\bar{x}_2)+\tau_2(\bar{x}_1))} \Big(-C + \left[\int_0^{\infty} e^{-\beta t} h_2(X_2(t))\, dt \;\middle|\; X_2(\tau_1(\bar{x}_2) + \tau_2(\bar{x}_1)) = \bar{x}_2 \right] \Big) \Big)}_{c)}, \end{aligned} \tag{14.4}$$

where $\tau_1(\bar{x}_2)$ is the smallest time at which the process $X_2(\tau_1(\bar{x}_2))$ is equal to $\bar{x}_2$ (i.e. X_2 is on $\mathcal{SO}_{2\to 1}$) and $\tau_2(\bar{x}_1)$ is the smallest time at which the process $X_1(\tau_2(\bar{x}_1))$ is equal to $\bar{x}_1$ (i.e. X_1 is on $\mathcal{SO}_{1\to 2}$). The term a) describes the reward received by engaging the project 2 until time $\tau_1(\bar{x}_2)$. The term b) describes the reward received when we switch from project 2 to project 1 at time $\tau_1(\bar{x}_2)$ and then we engage it until time $\tau_2(\bar{x}_1)$. The term c) describes the reward received when we switch from project 1 to project 2 at time $\tau_2(\bar{x}_1)$ and then we engage it forever.

Here again by definition of the switching curve, the value of $\bar{x}_2$ must satisfies:

$$\frac{\partial}{\partial \bar{x}_2} JO(x_1, v_1, 2; \bar{x}_2) = 0.$$

The switching curve $\mathcal{SO}_{1\to 2}$ on the interval $[u_1, A_1]$ is again given by symmetry. Iteratively, we clearly can compute the complete curve $\mathcal{SO}_{1\to 2}$.

Remark: For non-symmetric two-armed DMABP, the above procedure can be generalized straightforwardly. Indeed, the symmetry assumption is not required to iterate the construction of $\mathcal{SO}_{2\to 1}$.

14.3 ⊛ Explicitly Solved Example - Deteriorating and Deterministic MABP

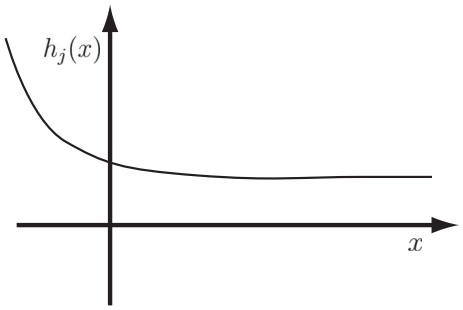

Fig. 14.5. Shape of the reward of class $\mathcal{D}$ of MABP.

To illustrate our method, let us compute explicitly the recursion for the following deterministic two-armed symmetric DMABP :

Consider the deterministic class of two-armed Bandit problems for which the dynamics $X_j(t)$ of project $j \in \{1, 2\}$ and its reward $h_j(x_1)$ read respectively as follows (see Figure 14.5):

$$\frac{dX_j}{dt} = \theta_j \quad ; \quad X_j(0) = x_0 \tag{14.5}$$

and

$$h_j(x) := \Gamma(1 + e^{-\alpha x}), \tag{14.6}$$

where $\alpha, \theta_j \in \mathbb{R}_+$ and $\Gamma \in \mathbb{R}$. This class of two-armed DMABP will be called as class $\mathcal{D}$ in the following. Note that $h_j(X_j(t_1)) < h_j(X_j(t_2))$, $\forall\ t_2 > t_1$, so that, the class $\mathcal{D}$ belongs to the class of DMABP (see Lemma 5.3).

For a Bandit in class $\mathcal{D}$, equation (14.2) reduces to:

$$\int_0^\infty e^{-\beta t}\Gamma(1 + e^{-\alpha(\theta_1 t + \infty)})dt = -C + \int_0^\infty e^{-\beta t}\Gamma(1 + e^{-\alpha(\theta_2 t + A_2)})dt,$$

from which we obtain:

$$A_2 = -\frac{1}{\alpha}\ln\left[\frac{C(\beta + \alpha\theta_2)}{\Gamma}\right].$$

The equation (14.3) reduces to:

$$JO(x_1, A_2, 2, \bar{x}_2) =$$

$$\int_0^{\tau(\bar{x}_2)} e^{-\beta t}\Gamma(1 + e^{-\alpha(\theta_2 t + A_2)})dt + e^{-\beta\tau(\bar{x}_2)}\left(-C + \int_0^\infty e^{-\beta t}\Gamma(1 + e^{-\alpha(\theta_1 t + x_1)})dt\right),$$

with

$$\tau(\bar{x}_2) = \frac{\bar{x}_2 - A_2}{\theta_2}.$$

equation (14.4) reduces to:

$$JO(x_1, v_1, 2, \bar{x}_2) = \int_0^{\tau_1(\bar{x}_2)} e^{-\beta t} \Gamma(1 + e^{-\alpha(\theta_2 t + v_1)}) dt +$$

$$e^{-\beta \tau_1(\bar{x}_2)} \Big(-C + \int_0^{\tau_2(\bar{x}_1)} e^{-\beta t} \Gamma(1 + e^{-\alpha(\theta_1 t + x_1)}) dt +$$

$$e^{-\beta(\tau_1(\bar{x}_2) + \tau_2(\bar{x}_1))} \left(-C + \int_0^{\infty} e^{-\beta t} \Gamma(1 + e^{-\alpha(\theta_2 t + \bar{x}_2)}) dt \right) \Big),$$

with

$$\tau_1(\bar{x}_2) = \frac{\bar{x}_2 - v_1}{\theta_2} \quad \text{and} \quad \tau_2(\bar{x}_1) = \frac{\bar{x}_1 - x_1}{\theta_1}.$$

These equations are transcendant for general values of α, β, θ_i, $i = 1, 2$. When $\alpha = \beta = \theta_1 = \theta_2 = 1$, explicit solutions can however be found and read:

$$A_1 = A_2 = -\ln\left[\frac{2C}{\Gamma}\right],$$

$$u_1 = v_1 = -\ln\left[\frac{6C}{\Gamma}\right],$$

$$u_2 = v_2 = -\ln\left[\frac{16C}{7\Gamma - \sqrt{33}\Gamma}\right],$$

$$\bar{x}_1 = -\ln\left[\frac{e^{-\bar{x}_2}}{2} - \frac{C}{\Gamma}\right].$$

Hence the switching curves for positive initial conditions $(X_1(0), X_2(0)) \in \mathbb{R}_+ \times \mathbb{R}_+$ are:

$$\mathcal{SO}_{2\to 1} = \begin{cases} \infty & \text{if } x_1 > A_1, \\ -\ln\left[\frac{e^{-x_1}}{2} - \frac{C}{\Gamma}\right] & \text{if } u_1 \le x_1 < A_1, \\ -\ln\left[\frac{2(\Gamma - e^{x_1} C)^2}{\Gamma e^{x_1}(2\Gamma + 2e^{x_1} C + \sqrt{\Gamma^2 + 14\Gamma C e^{x_1} + Q^2 e^{2x_1}}}\right] & \text{if } u_2 \le x_1 < u_1, \\ \vdots & \end{cases} \tag{14.7}$$

and

$$\mathcal{SO}_{1\to 2} = \begin{cases} -\ln\left[2\left(e^{-x_1} + \frac{C}{\Gamma}\right)\right] & \text{if } x_1 \ge A_1, \\ -\ln\left[\frac{2\Gamma + 2Ce^{x_1} + \sqrt{\Gamma^2 + 16\Gamma C e^{x_1}}}{2\Gamma e^{x_1}}\right] & \text{if } u_1 \le x_1 < A_1. \\ \vdots & \end{cases} \tag{14.8}$$

The above results are drawn in Fig. 14.1.

15

⊛ Explicit Expression for the GIH Indices - Deteriorating and Deterministic Two-Armed MABP

For the class $\mathcal{D}$ of DMABP given by equations (14.5) and (14.6), the optimal stopping time τ^* for problem $\tilde{\mathcal{P}}_j$ when the DM is initially engaged on project $\mathcal{T}$ reads as:

$$\tau^* = \begin{cases} 0 & \text{if } \gamma \geq \Gamma(1+e^{-x_0\alpha}) + C\beta \\ -\frac{x_0\alpha + \ln\left[-\frac{\Gamma+C\beta-\gamma}{\Gamma}\right]}{\alpha\theta_1} & \text{if } \Gamma + C\beta < \gamma < \Gamma(1+e^{-x_0\alpha}) + C\beta \\ \infty & \text{if } \gamma \leq \Gamma + C\beta \end{cases} . \tag{15.1}$$

To compute the switching index $\nu s_j(X_j(0))$ we solve equation (12.17) with the stopping time of the optimal policy τ^* given by equation (15.1) and with the identification:

$$\gamma = \nu s_j(X_j(0)).$$

This equation is generally transcendant. For the special case $\alpha = \beta = \theta_1 = \theta_2 = 1$, a closed form solution exists and reads as:

$$\nu s_1(x_0) = \Gamma(1+e^{-x_0}) + C - 2\sqrt{\Gamma C}\, e^{-\frac{x_0}{2}}, \tag{15.2}$$

Using this expression, we can explicitly characterize the switching curve resulting from the GIH of our symmetric two-armed DMABP. Indeed, we have:

$$S_{1\to 2} = \left\{ (x_1, x_2) \in \mathbb{R}^2 \mid \nu c_1(x_1) = \nu s_2(x_2) \right\} \Rightarrow$$

$$S_{1\to 2} = \left\{ (x_1, x_2) \in \mathbb{R}^2 \mid x_2 = -2\ln\left[e^{-\frac{x_1}{2}} + \frac{C}{\sqrt{\Gamma C}} \right] \right\} \tag{15.3}$$

and

$$S_{2\to 1} = \left\{ (x_1, x_2) \in \mathbb{R}^2 \mid \nu c_2(x_1) = \nu s_1(x_2) \right\} \Rightarrow$$

$$S_{2\to 1} = \left\{ (x_1, x_2) \in \mathbb{R}^2 \;\middle|\; \begin{cases} x_2 = -2\ln\left[e^{-\frac{x_1}{2}} - \frac{C}{\sqrt{\Gamma C}} \right] & \text{if } x_1 < -2\ln\left[\frac{\sqrt{\Gamma C}}{C}\right] \\ +\infty & \text{otherwise} \end{cases} \right\}. \tag{15.4}$$

In Fig. 15.1 we plot simultaneously the optimal hysteretic policy given by equations (14.7) and (14.8) and the GIH given by equations (15.3) and (15.4). This picture clearly shows that the optimal policy has a wider hysteretic gap. This behaviour is in agreement with the result expressed by Lemma 2.7 in [2] (see also Lemma 12.13 above).

Remarks:

- The claim and its demonstration can be generalized for DMABP when the dynamics of the project is given by random walks with no downward jumps (see [26] for the details).

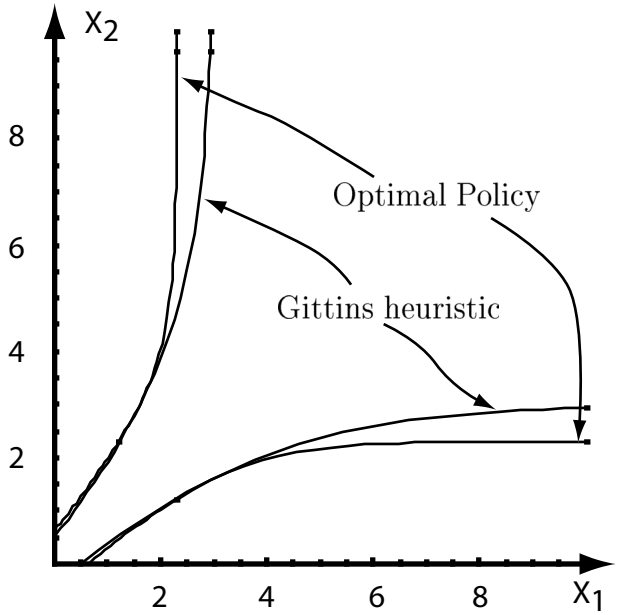

Fig. 15.1. Optimal policy and the GIH for the parameter values: $\alpha = \beta = \theta_i = 1$, $\Gamma = 2$, $C = 0.1$.

- The sub-optimality of the GIH can be observed by the explicit computation of the discounted reward obtained under a special initial condition. For example, choose $\Gamma = 2$, $C = 1.1$, $\alpha = \beta = \theta_1 = \theta_2 = 1$, and the initial conditions $(X_1(0) = 0; X_2(0) = 0; 1)$. With these values, the GIH commands to engage project 1 until the system reaches the position $(-2\ln\left[1 - \frac{C}{\sqrt{\Gamma C}}\right], 0)$, then to switch to project 2 and engage it forever. This scheduling yields a global reward of 2,988. Instead, the optimal policy commands to engage project 1 forever and yields a global reward of 3.

- For large values of β, the reward gained in the near future is dominant. Hence, when β is large enough, the reward realized after the first switching tends to be negligible and the GIH is expected to bring results closer to the optimal one. We observe this fact for the class of symmetric Bandits given by equation (14.6) by computing numerically the value $A_2 \equiv A_1$ and comparing it with the optimal one. Both values converge as β increases. A numerical example is given in the following table where we compute A_2 for $C = 0.1$, $\theta_i = \alpha = 1$, $\Gamma = 2$ and for three different values of β

β	A_2 GIH	A_2 optimal
1	2.996	2.302
5	1.386	1.203
10	0.571	0.597

16

Summary

We have seen in this part that the introduction of switching penalties in sequential decision problems drastically complicates them. Indeed, although the optimal policy for the MABP without switching penalties is known explicitly for the general cases, the introduction of switching penalties in MABP preclude the possibility to derive it generically. In particular, this optimal policy is no a naive extension of the Gittins index (see section 12.1 or [3]). Unfortunately, numerically solving the MABP with switching penalties is hopeless as the problems are enormously time and space consuming. Trying to approach the optimal policy by considering problems which do not take into account the switching penalties, give catastrophic results as they allow chattering (i.e. switching from one type of production to another arbitrarily often). However, as numerous optimization problems possess switching costs and/or time delays, we propose in chapter 12 a simple heuristic scheduling derived from a set of generalized priority indices. We show that our heuristic is more natural than the one first introduced in a contribution of M. Asawa and D. Teneketzis [2], in the sense that it follows directly from the Gittins index policy. Moreover, our heuristic gives, contrarily to the one of Asawa et al. the optimal policy for the class of

- MABP with switching penalties defined by problem $\tilde{\mathcal{P}}_j$ (see definition 12.3,

We then derive the optimal policy for the dynamic scheduling of a class of deterministic, deteriorating, continuous time and continuous state two-armed Bandit problems with switching costs. Due to the presence of switching costs, the scheduling policy exhibits an hysteretic character. We show that in presence of switching penalties, our heuristic also exhibits a hysteretic behaviour and reproduces the result already derived by Banks et al. [3] or Asawa et al. [2] for the switching costs problem. Finally, using this exactly solvable class of models, we are able to explicitly observe the performance of the proposed heuristic. It is observed that, on the studied example, the GIH yields result which are close to the optimal one. In particular the global shape of the optimal switching curves are given by the GIH.

Part IV

Dynamic Scheduling of a Flexible Machine.

17

Multi-Items Production Facility Operating on a Make-to-stock Basis

Introduction

In the present part, we explore the scheduling rules and the hedging levels that can be obtained by using a Restless Bandit Problem formulation of a make-to-stock production. In order to allow an analytical study we consider a special configuration of the production line where a single machine is able to manufacture N-types of items but only one at a time. This model is defined in section 17.1 and is translated to the Restless Bandit formalism in section 17.2. In order to be able to directly apply the result obtained in the previous part, the holding and backorder costs are supposed to be piecewise linear, and the stochastic processes describing the production and the demand flows are supposed to be Markov chains in continuous time or Diffusion processes. The RBP affords to analytically construct dynamic scheduling rules. These analytical results are compared in section 17.3 with the numerically derived optimal policy, obtained for a server delivering two types of items. It is observed that the Whittle relaxed version of the Restless Bandit model yields nearly optimal dynamic scheduling rules.

17.1 Flexible Manufacturing System - Definition

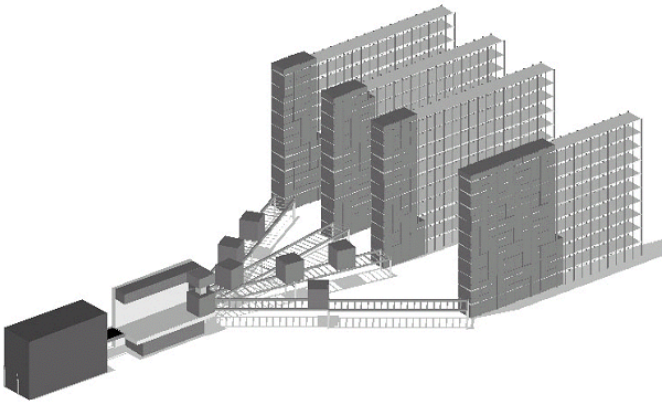

Fig. 17.1. Multiclass, make-to-stock production system.

The dynamic scheduling of tasks in flexible manufacturing systems belongs to the class of optimal decision problems. In its most general version, we are in presence of a flexible

workshop able to produce N different types of items, satisfying an external demand subject to fluctuations. We normally suppose that the flexible workshops have a limited capacity in the sense that they can produce only $K < N$ items at the same time. In order to serve the customers with the smallest delivery delays, the industrial approach is usually to produce on a make-to-stock basis (i.e. they build stocks of finished goods). Obviously there is a tradeoff between the storage costs and the costs incurred by late service of the customers demands. Therefore, the characterizations of the optimal

i) hedging stock capacities

and the

ii) scheduling rules which select the type of items to produce at a given time,

is the core of the production problems. Clearly, a small hedging level risks to incur lost sales penalties while a large one implies expensive storage costs.

This problem becomes even more complex when the switching between one type of production to another generate additional costs (the need of an additional workforce for example) and switching time delays (a cleaning operation for example). The general problem of a flexible machine having switching penalties was too general to be treated as part of this thesis but it is the natural continuation of this work. In the following, we shall therefore focus our attention to problems for which the **set-up costs and/or time delays are negligible**. Let us now give a precise definition of the flexible machine problem in order to answer both questions *i*) and *ii*) above.

The problem of a multiclass, make-to-stock production system subject to breakdowns and repairs is composed of a machine able to produce N different types of items. Finished items are stored in N different respective finished good inventories (FGI) (see Figure 17.1). We write

$$\vec{X}(t) = (X_1(t), \ldots, X_j(t), \ldots, X_N(t)) \in \Omega^N \subset \mathbb{R}^N \tag{17.1}$$

to describe the net inventory process characterizing the inventory levels in the FGI's at time t. Let us denote by $0 \leq d_j < \infty$ the capacity of the stock of item of type j. We then have:

$$\Omega^N =]-\infty, d_1[\times \cdots \times]-\infty, d_N[\subset \mathbb{R}^N$$

The negative values of $X_j(t)$ describe the presence of backorders (i.e. demands that cannot be immediately satisfied). Let us assume that we can schedule the production facility by using a "bang-bang" type control variable

$$\vec{u}(t) = (u_1(t), \ldots, u_N(t)) \ ; \quad u_j(t) \in \{0, 1\}$$

where $u_j(t) = 1$ means that the production of the items of type j is engaged and $u_j(t) = 0$ means that no item of type j is produced. Note that $u_j(t) = 1$ only implies that the production of the items of type j is engaged but it does not imply that the machine is actually producing an item of type j, indeed the machine can be either broken down or the stock can be filled up so the machine is blocked.

Both the instantaneous controlled production rate $\vec{P}(t)$ and demand rate $\vec{D}(t)$ are supposed to be independent, stationary Markovian random vector-processes. We shall respectively write:

$$\vec{P}(t) = \big(P_1(t), \ldots, P_j(t), \ldots, P_N(t)\big) \tag{17.2}$$

and

$$\vec{D}(t) = \big(D_1(t), \ldots, D_j(t), \ldots, D_N(t)\big). \tag{17.3}$$

The $P_j(t)$ are independent random processes which model the controlled production rates for the items of type j and the $D_j(t)$ are independent random processes which model the demands for the items of type j. Moreover, the processes $P_j(t)$ satisfy the constraints:

$$0 \leq P_j(t) \leq \mathcal{C}_j(t),\ t \geq 0,\ j = 1, \ldots N,$$

where $\mathcal{C}_j(t) \in \{0, c_j\}$ denote the stochastic nature of the production capacity of the (failure-prone) machine. Here we assume $\mathcal{C}_j(t)$ to be alternating Markov renewal processes (i.e. two-states Markov processes). For an item of type j, the value $\mathcal{C}_j(t) = c_j$ is its maximal instantaneous production rate and $\mathcal{C}_j(t) = 0$ represents the failure state of the machine (see Figure 17.2).

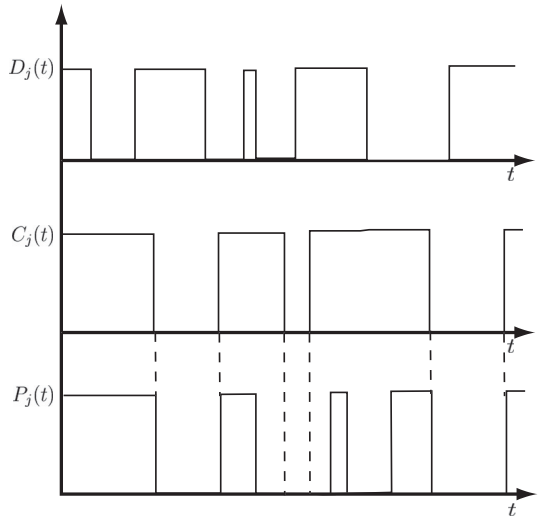

Fig. 17.2. Problem $\mathcal{P}_j$.

We impose the following condition on the controlled production process $\vec{P}(t)$:

a) The production facility has a limited capacity. Accordingly, it can engage at each time the production of a single item (i.e. at each time $t \in \mathbb{R}^+$, at most one of the N controls $u_j(t)$ is equal to one).

b) The machine must complete the production of the item currently in process before starting a new one. For a single server with a capacity limited to one type, this assumption is natural.

To pilot the production, we introduce a scheduling policy π, which is a mapping:

$$\begin{aligned} \pi \ :\ \Omega^N \ &\to \{0,1\}^N \\ \vec{X}(t) &\mapsto \ \vec{u}(t). \end{aligned}$$

The function $\vec{u}(t)$ defines which type of item the machine is manufacturing at time t.

Definition 17.1 (Admissible policy). *A policy $\pi(t)$ is admissible if it depends only on the present state of the random net inventory process $\vec{X}(t)$.*

In the sequel, we shall only consider admissible policies for which $\vec{P}(t)$ and $\vec{u}(t)$ fulfill both conditions a) and b) above. The set of admissible policies will be called as $\mathcal{U}$.

Given a policy $\pi \in \mathcal{U}$, let us denote the cumulative production up to time t by

$$\vec{\mathcal{P}}_\pi(t) = \int_0^t \vec{P}(t)\, dt$$

and the cumulative demand up to time t by

$$\vec{\mathcal{D}}(t) = \int_0^t \vec{D}(t)\, dt.$$

In terms of $\vec{\mathcal{P}}_\pi(t)$ and $\vec{\mathcal{D}}(t)$, the net inventory process can therefore be expressed as:

$$\vec{X}(t) = \vec{\mathcal{P}}_\pi(t) - \vec{\mathcal{D}}(t), \tag{17.4}$$

with the initial condition

$$\vec{X}(0) = \vec{x}_0 \in \Omega^N.$$

Let us further introduce the instantaneous running costs $h(\vec{X}(t), \vec{u}(t))$ depending on the global state of the net inventory $\vec{X}(t)$ and the production engagement $\vec{u}(t)$:

$$h \;:\; \Omega^N \times \{0,1\}^N \;\rightarrow\; \mathbb{R}^N$$

$$(\vec{X}(t), \vec{u}(t)) \;\mapsto\; h(\vec{X}(t), \vec{u}(t)) = (h_1^{\theta_1}(x_1), \ldots, h_N^{\theta_N}(x_N)),$$

where $\theta_j = a$ (active) when the production of the j-type items is engaged (i.e. $u_j(t) = 1$) for a stock level $X_j(t) = x_j$ and $\theta_j = p$ (passive) when the production is switched off (i.e. $u_j(t) = 0$). In the following, the instantaneous cost will be the storage cost and cost incurred by late delivery to a customer. The storage cost of item of type j only depends on the quantity of finished goods present in the stock X_j and not on the production engagement $u_j(t)$ of the flexible machine. It is reasonable to assume that the storage cost is linear with respect to the number of finished goods present in the stocks. Similarity, the cost incurred by late delivery only depends on the number of customers waiting to be served and not on the production engagement. It is also reasonable to assume that this cost is linear with respect to the number of customers waiting. Nevertheless, the cost incurred by late delivery is generally greater than the cost incurred by storage. Hence, we will assume that

$$h_j^a(x) \equiv h_j^p(x) =: h_j(x)$$

where $h_j(x)$ is a piecewise linear function.

Assuming an infinite time horizon, we define for an initial condition $\vec{x}_0$ the total discounted production cost $J^\pi(\vec{x}_0)$ under policy π by:

$$J^\pi(\vec{x}_0) := E^\pi_{\vec{x}_0} \int_0^\infty e^{-\beta t} h(\vec{X}(t), \vec{u}(t))\, dt, \tag{17.5}$$

with $E^\pi_{\vec{x}_0}$ denoting the expectation operator, conditioned on $\vec{x}_0$ and where $e^{-\beta t}$ with $\beta > 0$ is a discounting factor. This discounting factor can be understood as the running interest on the stocked raw material and finished goods, plus the running interest on the location of the production facility (the machines, the rooms...). Clearly, the integral in equation (17.5) does not exist for arbitrary cost functions $h(\vec{x}, \vec{u})$. In the following, we assume that $h(\vec{x}, \vec{u})$ satisfies the required regularity conditions, to insure that equation (17.5) exists $\forall\ \pi \in \mathcal{U}$ and for all initial conditions.

The optimal control problem is to determine the optimal policy π^* minimizing the total production cost given in equation (17.5). The value of the production cost under the optimal policy π^* will be written as:

$$J^*(\vec{x}_0) := J^{\pi^*}(\vec{x}_0) = \inf_{\pi \in \mathcal{U}} J^{\pi}(\vec{x}_0).$$

In absence of setup penalties, we will see that the formalization given by the Restless Bandit Problem (RBP) is relevant to model the scheduling of flexible production systems. Finding the optimal scheduling rule for a flexible production then essentially amounts to finding the optimal policy for the RBP. In chapter 8, we have seen that using the Bandit formalization enables us to approximately decouple the original N-armed RBP into N single-item make-to-stock production processes [50]:

$$X_j(t) = \mathcal{P}_j(t) - \mathcal{D}_j(t), \ j = 1, \ldots, N, \tag{17.6}$$

for which the scheduling policy is much easier to determine. In fact, the decoupling into single production processes enables us to construct Priority Indices $\nu_j(x_j)$ used to construct a reliable heuristic based on the "Priority Index Policy".

Remark: For the scheduling production problem, we will introduce an index $\nu_{N+1}(x)$ taking into account the fact that the machine can stay idle. In the following, we will call this index "*the idle index*". When this idle index is smaller than all the other indices, the machine does not produce any item.

Remember that priority index policies give optimal rules for the Multi Armed Bandit Problem (MABP), characterized by the fact that if not engaged, the projects remain "frozen" (i.e $X_j(t) = X_j(t + dt)$) and no cost is incurred (i.e. $h_j^p(x) \equiv 0$). This former assumption is however not fulfilled for the multi-class production systems we are dealing with. Indeed, even unserved the demands continue to increases and hence the global state of the system is in permanent evolution (i.e. the net inventory of an item not currently produced evolves with time). In this case, the scheduling of a flexible production naturally belongs to the class of the RBP. For the general N-item scheduling problem, similar decoupling approximation methods form the core of several recent contributions where priority index policies are shown to provide suboptimal, but efficient, scheduling rules ([15], [1], [35]). In particular, in [35], the authors use, as in the present section, the RBP to describe an heuristic scheduling rule for a multiclass, make-to-stock $M/M/1$ queuing system. The Whittle index for the case where the machine operates with lost sales and under the time average cost criterion (i.e. $\beta = 0$), is explicitly computed. Our contribution, published in [12], is complementary as we explicitly compute the Whittle index for machines working **with backorder** and under the **discounted cost criterion** (i.e. $\beta > 0$).

17.2 ⊛ Dynamic Scheduling of Multiclass Make-to-Stock Production.

The make-to-stock problem can be naturally formulated as a multi-armed RBP (defined in part II) with $N+1$ projects as follows:

- Identify the positions of the N net inventories $X_j(t)$, $j = 1, 2, \ldots N$, with the first N projects of the RBP.
- Take the active and the passive cost functions identical for each item (i.e. $h_j^a(x) = h_j^p(x) =: h_j(x)$).
- Add an extra project $X_{N+1}(t)$, called the *idle* project, with "frozen" dynamics given by:
$$X_{N+1}(t) \equiv \xi, \ t \in \mathbb{R}^+$$
i.e. the idle project take a constant value ξ and engaging the idle project models the decision to be idle.
- Impose the idle project to incur no cost (i.e. $h_{N+1}(x) \equiv 0$).

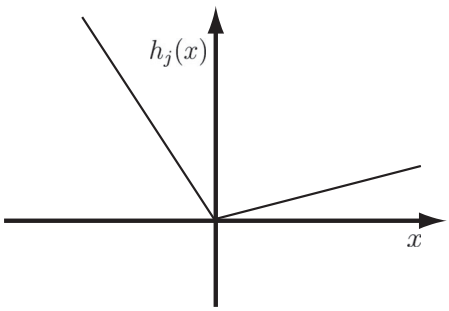

Fig. 17.3. Piecewise linear costs function.

In order to be able to perform explicit computation, we assume the cost functions $h_j(x)$ to have a piecewise linear form (see Figure 17.3):

$$h_j(x) = A_j x^+ + B_j x^-, \quad j = 1, 2, \ldots N, \tag{17.7}$$

where $x^+ = \max(x, 0)$; $x^- = \max(-x, 0)$ and $A_j, \ B_j \geq 0$.

By construction, the index of the idle project is $\nu_{N+1}(x_c) \equiv 0$. Therefore, following the Whittle heuristic, the machine is left idle (i.e. does not produce any item) when all indices are positive (i.e. $\nu_j(x_j) > 0$, $j = 1, \ldots, N$). Moreover, as all indices are strictly increasing, the positions $d_j \in \mathbb{R}$ such that $\nu_j(d_j) = 0$ correspond to the levels of the hedging stocks for the items of type j.

17.2.1 ⊛ Markovian Queue Dynamics

Assume that the net inventory process given by equation (17.4) is described by a continuous-time discrete state Markov chain with parameters λ_j and μ_j. The resulting make-to-stock problem is identical to the one studied in [22]. Note also that the contribution [35]

discusses a similar problem, but the optimization is done under the average cost criterion.

Following [35], we apply the standard uniformization argument given in [30] and the Dynamic Programming equation (8.4) becomes:

$$J^*(\vec{X}) = \frac{1}{\Lambda+\beta}\left[h(\vec{X}) + \sum_{j=1}^{N}\lambda_j J^*(\vec{X}-e_j) + \mu J^*(\vec{X}) + \min\left\{0, \min_j\left(\mu_j\beta_j J^*(\vec{X})\right)\right\}\right], \tag{17.8}$$

with $\beta_j J^*(\vec{X}) = J^*(\vec{X}+e_j) - J^*(\vec{X})$ and e_j is the unit vector with the j-th component equal to unity,

$$\mu = \max_j\{\mu_j\} \quad \text{and} \quad \Lambda = \mu + \sum_{i=1}^{N}\lambda_i.$$

As stated in [35], the form of equation (17.8) suggests that the optimal policy can be described with switching curves and hedging stocks (this is indeed proven in [22]). As the priority index policies lead to a similar structure, the Whittle relaxation method is suitable to discuss the production problems. Note that strictly speaking, the Priority Index Policy cannot be used directly to construct a heuristic from the Whittle index in general. Indeed, in the flexible manufacturing problem, the machine has to complete the currently engaged item before starting a new one. On the other hand, the policy resulting from the Whittle relaxation is based on indices defined on $\mathbb{R}$. Therefore, a direct use of the multi-armed RBP requires a change in the production before completing the engaged item. To overcome this difficulty and then allow the use of the RBP in this production context, we will suitably renormalize the service time in order to approximately take into account the time necessary to complete engaged items. The renormalization process is achieved by imposing:

$$\mu_j \mapsto \tilde{\mu}_j = \rho_j\mu_j + (\rho-\rho_j)\frac{1}{\frac{1}{\mu_j}+T} + (1-\rho)\mu_j, \tag{17.9}$$

$$\rho_j = \frac{\lambda_j}{\mu_j}, \quad \rho = \sum_{j=1}^{N}\rho_j.$$

In writing equation (17.9) we have used the fact that when the production priority (derived from the Whittle heuristic) is to engage the type j production, three alternatives may occur, namely:

a) The server is already engaged on the type j products and the average service time is $\frac{1}{\mu_j}$.

b) The server is engaged on a production of type $k \neq j$ and the average service time is $\frac{1}{\mu_j} + T_k$ where T_k is the average time needed to finish the production of the type k item. We denote by T the average of the T_k.

c) The server is idle and, as we consider only problems without switching time, the average service time is $\frac{1}{\mu_j}$.

For infinite time horizon, $\tilde{\mu}_j$ is therefore a weighted average taking into account the relative contributions of a), b) and c). The respective weights are determined as follows:

i) The average sojourn time in situation a) is proportional to the partial traffic $\rho_j = \frac{\lambda_j}{\mu_j}$.

ii) The average sojourn time in situation b) is proportional to the global traffic, minus the partial traffic of the k-type items: $(\rho - \rho_k)$.

iii) The average sojourn time in situation c) is proportional to the percentage of idle time: $(1 - \rho)$.

Note that the value of T is bounded:

$$0 \leq T \leq \max_{j \in \{1,\ldots,N\}} \left(\frac{1}{\mu_j}\right) = \frac{1}{\mu}$$

Remark: The renormalization procedure given by equation (17.9) is not necessary when the ratio $\frac{\beta}{\mu}$ is large. Indeed, in this case, the characteristic discount time $\frac{1}{\beta}$ being much smaller than the characteristic switching time delays $\frac{1}{\mu}$ (i.e. $\frac{1}{\beta} << \frac{1}{\mu}$), the costs incurred after the first few decisions tend to be negligible. Hence the global cost differences induced by changing before or after the completion time will be negligible when $\frac{\beta}{\mu} >> 1$.

Let us now apply the Whittle relaxation and hence focus on a single item problem, say item j. Then the γ-penalty given by problem (8.3) reads as:

$$J^j(x,\gamma) = \frac{1}{\lambda_j + \mu_j + \beta}\left[h_j(x) + \lambda_j J^j(x-1) + \mu_j J^j(x) + \min\{\gamma, \mu_j \beta J^j(x)\}\right]. \quad (17.10)$$

From now on, we suppress the index j as the computation involves only a single item. To make headway, we assume the indexability property to hold. Following the method of section 8.1.3, equation (8.4) reads as:

$$\begin{cases} \beta J_a(x,\gamma) = h(x) + \lambda J_a(x-1,\gamma) + \mu J_a(x+1,\gamma) - (\lambda+\mu) J_a(x,\gamma) \\ \beta J_p(x,\gamma) = h(x) + \lambda J_p(x-1,\gamma) - \lambda J_p(x,\gamma) + \gamma \end{cases} \quad (17.11)$$

which is a special case of the system given in equation (8.16) with $\mu_p = 0$.

From section 8.1.3, we have:

$$\begin{array}{l} J_a(x,\gamma) = C_a^+ (w_+)^x + C_a^- (w_-)^x + S_a(x,\gamma) \\ J_p(x,\gamma) = C_p (w_0)^x + S_p(x,\gamma), \end{array} \quad (17.12)$$

with C_a^+, C_a^-, C_p being integration constants,

$$\begin{array}{l} w_+ = \frac{(\beta+\lambda+\mu) + \sqrt{(\beta+\lambda+\mu)^2 - 4\lambda\mu}}{2\mu}, \\ w_- = \frac{(\beta+\lambda+\mu) - \sqrt{(\beta+\lambda+\mu)^2 - 4\lambda\mu}}{2\mu}, \\ w_0 = \frac{\lambda}{\lambda+\beta} \end{array} \quad (17.13)$$

and $S_a(x,\gamma)$, $S_p(x,\gamma)$ are the relevant particular solutions. Explicit computations are given in Appendix C, where we find:

$$S_a(x,\gamma) = \frac{1}{\mu(w_+ - w_-)}\left\{h(x) + (w_-)^x \sum_{j=-\infty}^{x-1} h(k)(w_-)^{-k} + (w_+)^x \sum_{j=x+1}^{\infty} h(k)(w_+)^{-k}\right\},$$

and

$$S_p(x,\gamma) = (w_0)^{x+1} \sum_{j=-\infty}^{x} \frac{(h(k)+\gamma)(w_0)^{-k}}{\lambda}.$$

Using the fact that $0 \leq w_- \leq w_0 \leq 1 \leq w_+$ and the asymptotic behaviour of the solutions, we have:

$$\lim_{x \to -\infty} J_a(x, \gamma) = \lim_{x \to -\infty} S_a(x, \gamma) \Rightarrow C_a^- = 0.$$

As before, $\nu(x)$ is computed with the smooth-fit principle expressed in equation (8.8). For the piecewise linear cost function $h(x)$ given by equation (17.7), the corresponding index, derived in Appendix D, reads (see Figure 17.4):

$$\nu(x) = \begin{cases} \frac{A\mu - \mu(A+B)(w_-)^{x+1}}{\beta} & \text{if } x \geq 0 \\ -\frac{B\mu}{\beta} & \text{if } x < 0. \end{cases} \tag{17.14}$$

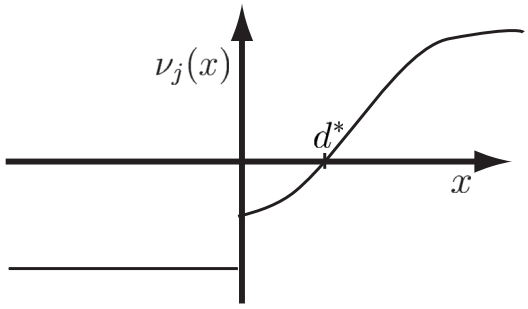

Fig. 17.4. Typical shape of the index for Markovian queue dynamics.

Remarks:

- Observe that with our choice of $h(x)$ the index $\nu(x)$ is monotonically increasing. This is a necessary property for indexability ([36] and [37]). Hence the Whittle heuristic for the single item problem coincides with a hedging stock policy (i.e. *produce only when the stock level is below the hedging stock*), with a hedging level d^* defined as:

$$\nu(d^*) = 0. \tag{17.15}$$

  The hedging stock policy is known to be optimal for a single-item production [35] and [6]. Solving equation (17.15) with equation (17.14), we find:

$$d^* = \max\left\{0, \left\lceil \ln\left(\frac{A}{A+B}\right)\frac{1}{\ln(w_-)} - 1 \right\rceil\right\}, \tag{17.16}$$

  where $\lceil x \rceil = n$ with $n \in \mathbb{Z}$ is the smallest integer value larger than x. Despite to the fact that the extended smooth-fit principle (see equation (8.8)) used to derive equation (17.15) does not yield the optimal solution in general, here it does. Indeed, the optimal hedging level d^* can be derived by using an alternative approach explained in [20]. This computation performed in Appendix B yields:

$$d^* = \max\left\{0, \left\lfloor \ln\left(\frac{A}{A+B}\right)\frac{1}{\ln(w_-)} \right\rfloor\right\}, \tag{17.17}$$

  where $\lfloor x \rfloor = n$ with $n \in \mathbb{Z}$ is the largest integer value smaller than x. Clearly both hedging stocks equations (17.16) and (17.17) are identical.

- Asymptotic regimes: In the limit $\beta \to 0$, we have that $w_- \to \rho := \frac{\lambda}{\mu}$ and therefore:

$$\lim_{\beta \to 0} d^* = \max \left\{ 0, \left\lfloor \frac{1}{\ln(\rho)} \ln \left(\frac{A}{A+B} \right) \right\rfloor \right\}.$$

  This is the optimal hedging point for the single-item problem under the average cost criterion, derived in [35]. Moreover, when $\beta \to 0$ and $\rho \to 1$ (heavy traffic limit), the optimal hedging stock tends to infinity (i.e. $d^* \to \infty$). Such a limiting behaviour also follows from equation (3.7) of [14], when $c \to 0$ and $\gamma \to 0$.

- For a multi-item problem, the Priority Index Policy solves the scheduling production problem only sub-optimally in general. This will be explicitly seen in the numerical simulation exposed in section 17.3.

- As it is emphasized in [35], the index equation (17.14) does not exist in the limit $\beta \to 0$ (i.e. $\nu(x) = -\infty$). On the contrary, for $\beta > 0$, the scheduling policy does not need to serve a fixed time-average number of classes. Accordingly, for the relaxed version of the RBP, priority indices exist when $\beta > 0$.

- When $x < 0$, the Restless priority index directly reduces to the $B\mu$ policy. This is consistent with the scheduling rule proposed in Wein [48] and in Ha [22]. Moreover, the asymptotic behaviour:

$$\nu(x) = \begin{cases} \frac{A\mu}{\beta} & \text{when } \; x \to +\infty, \\ \frac{-B\mu}{\beta} & \text{when } \; x < 0, \end{cases} \tag{17.18}$$

  exhibits the structure of the well-known $A\mu/B\mu$ policy (see definition 17.4 below for a more detailed discussion devoted to this policy).

- When the indexability property is fulfilled, the make-to-stock single item production problem is optimally solved by the Whittle relaxation. This can be proved as follows:

  Consider the two-armed RBP formulation of the single item production problem and remember that the idle project does not incur cost. This two-armed RBP is equivalent to the γ-penalized problem (17.10) with $\gamma = 0$. The optimal policy for this problem is to engage the production in all states $x \in \mathbb{Z}$ for which $\nu(x) < \gamma = 0$. As the idle index is $\nu_{N+1}(x) \equiv 0$, the optimal policy for the $\gamma = 0$-penalized problem is therefore equivalent to engaging the project with the smallest priority index. Hence the Whittle relaxation indeed solves the single item production problem optimally.

- In [23], Dallery et al. consider the problem of minimizing the storage and backorder costs incurred by a flexible production facility having the following property:

  - There is a few type of items, written by A, for which the demand is comparatively large.
  - There is a large number of type of items, written B for which the demand is reduced.
  - The storage and backorder cost are piecewise linear.
  - The net inventory process is described by a continuous-time discrete state Markov chain.

Dallery et al. show that among other, a very efficient policy is to build finished goods inventories for the items in A and produce on a make-to-order base for the items in B (i.e. $d_j > 0$ for $j \in A$ and $d_j = 0$ for $j \in B$). Moreover, the demands for items in B are fulfilled with priority (i.e. the production of items in A is stop when demands for a item in B arrive).

Applying the RBP to this problem restitutes precisely the policy derived in [23]. Indeed, computing the hedging level d_j, $j \in A$ given by equation (17.16), gives $d_j > 0$ and computing d_k, $k \in B$ gives $d_k = 0$. Moreover, the value $\nu_k(x)$ for $x < 0$ and $k \in B$ (i.e. $\nu_k(x) = -\frac{B_k \mu}{\beta}$) is smaller than $\nu_j(x)$ for $x < 0$ and $j \in A$. Hence when a demand for B arrives, the Priority Index Policy command to stop the production of A and start the production of B.

17.2.2 ⊛ Comparison of the RBP Heuristic with the Optimal Policy Derived by de Véricourt, Karaesman and Dallery

In [15] the authors consider a two-items make-to-stock single machine with Poisson process dynamics for the production time and the inter-arrival demand time. They derive numerically the optimal dynamic scheduling minimizing the average inventory and back-order cost under the time average criterion (i.e. $\beta = 0$). As noted before, the use of RBP in the limit $\beta \to 0$ requires the use of the renormalisation given by equation (17.9). Hence, after renormalization we compare in Figure 17.5 the optimal result derived in [15] with the RBP heuristic. We draw two different representative configurations of their problem which data are displayed in Table I. Observe that the global structure of the scheduling policy is given by the RBP. Remark that this structure will not follow from the use of a simple (myopic) $h\mu/b\mu$ policy. In particular, the fact that $B > 0$ (in Figure 17.5) and the linearly growing behaviour of the switching curve cannot follow from the $h\mu/b\mu$ rule. Note finally that the use of the Whittle relaxation leads to an underestimation of the hedging levels. This reflects the fact that the decoupling between the projects which follows from the Whittle relaxation is only approximative. A probable explication of this underestimation may be the following. By decoupling the original problem, we indeed do not fully take into account the limited capacity of the production facility. Indeed, when delivering a specific type of items, the server has a very high capacity compared with the demand rate. hence low hedging levels will optimally be required. The renormalization procedure given by equation (17.9) does only partly correct this effect.

Case	λ_1	λ_2	μ_1	μ_2	A_1	A_2	B_1	B_2
1	0.4	0.4	1	1	1	1	50	25
2	0.4	0.4	1	1	1	1	50	5

Table I

17.2.3 ⊛ Diffusive Dynamics

Let us finally consider the case where we model the demand respectively the production by diffusive processes following the stochastic differential equations:

$$dP_j(u_j(t), t) = u_j(t)\left[\mathcal{U}_j dt + \sigma_{j,P} dW_{j,P}(t)\right], \ j = 1, 2, \ldots N \tag{17.19}$$

respectively

$$dD_j(t) = \mathcal{V}_j dt + \sigma_{j,D} dW_{j,D}(t), \ j = 1, 2, \ldots N, \tag{17.20}$$

where $dW_{j,P}(t)$ and $dW_{j,D}(t)$ are independent White Gaussian Noise processes, $\mathcal{U}_j$ and $\mathcal{V}_j$ are the drifts and $\sigma_{j,P}$ and $\sigma_{j,D}$ are the variances of the diffusion processes.

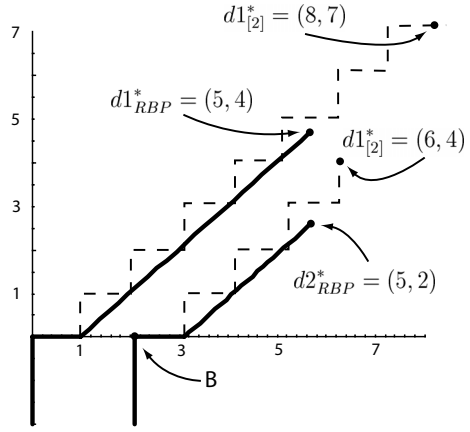

Fig. 17.5. Optimal switching curve derived in [15] (dashed line) and RBP heuristic (plain line).

Using equation (17.19) and equation (17.20), it is straightforward to write the time evolution of the net inventory $X_j(t)$ given by equation (17.4) as:

$$dX_j(t) = (\mathcal{U}_j u_j(t) - \mathcal{V}_j)\, dt + \sigma_j(u_j(t))dW_j(t), \tag{17.21}$$

where $dW_j(t)$ are standard independent WGN's for $j = 1, 2 \ldots N$, and the controlled variances $\sigma_j(u_j(t))$ read:

$$\sigma_j^2(u(t)) = (\sigma_{j,D})^2 + (u(t)\sigma_{j,P})^2\,,\ j = 1, 2 \ldots N. \tag{17.22}$$

Assume again that the running cost is piecewise linear, as in equation (17.7). Then for the Whittle relaxation problem, the value of the index $\nu_j(x_j)$ follows from equation (8.14), provided that the following relations hold:

$$\mu_{a,j} = (\mathcal{U}_j - \mathcal{V}_j) > 0,$$

$$\tilde{\mu}_{p,j} = -\mathcal{V}_j < 0,$$

$$\sigma^2_{a,j} = \sigma^2_j(u(t) = 1) = \sigma^2_{j,D} + \sigma^2_{j,P},$$

$$\sigma^2_{p,j} = \sigma^2_j(u(t) = 0) = \sigma^2_{j,D}.$$

To further simplify the analysis, we will assume that only the demand process $\vec{D}(t)$ fluctuates. Hence, we take $\sigma_{a,j} = \sigma_{p,j} = \sigma$. In this case, we can establish:

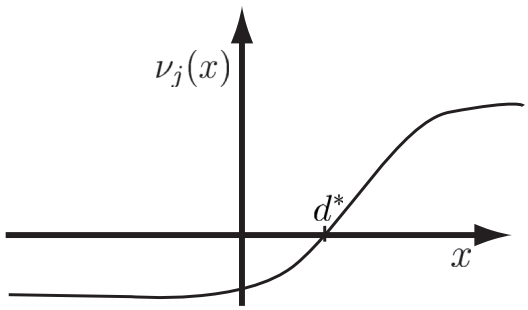

Fig. 17.6. Typical shape of the index for diffusive dynamics.

Lemma 17.2. *The index is monotonically increasing for every positive convex function $h_j(x)$ satisfying (see Figure 17.6):*

$$\int_0^\infty e^{-\beta t} h_j(\vec{X}(t))dt < \infty.$$

Proof: With the above assumptions, the index equation (8.14) simplifies and reads as (we drop the index j):

$$\nu(x) = \mathcal{K} \int_0^\infty e^{-y} \left[h\left(x + \frac{y}{w_a^+}\right) - h\left(x - \frac{y}{w_p^+}\right) \right] dy,$$

with

$$\mathcal{K} = \frac{w_p^+ - w_a^-}{w_a^-} > 0.$$

Hence:

$$\frac{d}{dx}\nu(x) = \mathcal{K} \int_0^\infty e^{-y} \left[\frac{d}{dx} h\left(x + \frac{y}{w_a^-}\right) - \frac{d}{dx} h\left(x - \frac{y}{w_p^+}\right) \right] dy.$$

This last expression is positive as the reward cost function is supposed to be convex ($\frac{d}{dx}h(x)$ is increasing). Then the index $\nu_j(x)$ is monotonically increasing .

□

Remark: The monotonously increasing nature of $\nu(x)$ is a necessary condition for indexability (see [36] and [37]).

Lemma 17.3. *When $B \geq A$, the hedging stock level d^* is given by*

$$d^* = \max\left\{0\,;\ \frac{\sigma^2}{\mu_a + \sqrt{\mu_a^2 + 2\beta\sigma^2}} \ln\left[\frac{(A+B)}{A}\left\{\frac{\mu_a + \mu_p - \sqrt{\mu_a^2 + 2\beta\sigma^2} + \sqrt{\mu_p^2 + 2\beta\sigma^2}}{2(\mu_a + \mu_p)}\right\}\right]\right\}. \tag{17.23}$$

Proof: For the cost function equation (17.7), the index can be explicitly written as:

If x is positive:

$$\nu(x) = \frac{1}{2\beta^2}\left(2A\left(\mu_a + \mu_p\right) + e^{-\frac{\left(\mu_a + \sqrt{\mu_a^2 + 2\beta\sigma^2}\right)x}{\sigma^2}}(A+B)\left[-\mu_a + \sqrt{\mu_a^2 + 2\beta\sigma^2} - \mu_p - \sqrt{\mu_p^2 + 2\beta\sigma^2}\right]\right). \tag{17.24}$$

If x is strictly negative:

$$\nu(x) = \frac{1}{2\beta^2}\left(-2B\left(\mu_a + \mu_p\right) + e^{\frac{\left(\mu_p + \sqrt{\mu_p^2 + 2\beta\sigma^2}\right)x}{\sigma^2}}(A+B)\left[+\mu_a + \sqrt{\mu_a^2 + 2\beta\sigma^2} + \mu_p - \sqrt{\mu_p^2 + 2\beta\sigma^2}\right]\right) \tag{17.25}$$

Now when $B \geq A$, solving $\nu(x) = 0$ with $\nu(x)$ given by equation (17.24), gives the required formula.

□

Remarks:

- When $B = 0$ and $A > 0$, the logarithm of the expression for d^* given in equation (17.23) is smaller than the unity and we consistently conclude that the optimal hedging is located at 0 (i.e."just-in-time" production-rule).

- In [14], Krichagina et al. derive for a regulated Brownian Motion a hedging level of the form

$$\hat{d}^* = \frac{\sigma^2}{\mu_a + \sqrt{\mu_a^2 + 2\beta\sigma^2}} \ln\left[\frac{A+B}{A}\right] \tag{17.26}$$

  Clearly equation (17.26) and equation (17.23) are identical up to the curly bracket factor present in equation (17.23). The difference can be traced back to the fact that in [14] a regulated Brownian Motion is assumed while here a "bang-bang" control is considered. These different control rules modify the local time of the process on the hedging level and hence the hedging point. In addition, note that for the limiting case $\beta \to 0$, equation (17.26) reduces to

$$\hat{d}^* = \frac{\sigma^2}{2\mu_a} \ln\left[\frac{A+B}{A}\right], \tag{17.27}$$

  which is the result derived by Wein [48]. This limit behaviour is expected as Wein studies an essentially similar situation to the one described in [14]. Indeed, Wein considers the scheduling under the average cost criterion ($\beta = 0$) of a multi-items facility with Markovian dynamics by using a diffusive limit.

  In the $\beta \to 0$ limit, our hedging stock equation (17.23) reduces to

$$d^* = \frac{\sigma^2}{2\mu_a} \ln\left[\frac{A+B}{A}\left\{\frac{\mu_p}{\mu_a+\mu_p}\right\}\right], \quad \frac{\mu_a}{\mu_p} \in [1, \infty[. \tag{17.28}$$

  The difference between equation (17.27) and equation (17.28) originates again from the different control rules on the hedging level.

  Note that in the limiting case $A = B$ and $\mu_a = \mu_p$, we immediately deduce that $d^* = 0$ from both equations (17.28) and (17.23). This is in perfect agreement with intuition, as for this symmetric case with a bang-bang control process it will clearly never be optimal to build a stock. This behaviour validates the hedging levels given by equations (17.28) and (17.23).

- As in section 17.2.1, if $B \neq 0$ and for a vanishing discounting factor $\beta \to 0$, the index $\nu(x)$ does not exist (i.e. tends to $-\infty$).

- Observe from equations (17.24) and (17.25) that we have the asymptotic behaviour:

$$\nu(x) = \begin{cases} \frac{A}{\beta^2}(\mu_a+\mu_p) = \frac{A}{\beta^2}\,\mathcal{U} & \text{when } \; x \to \infty \\ \\ \frac{-B}{\beta^2}(\mu_a+\mu_p) = \frac{-B}{\beta^2}\,\mathcal{U} & \text{when } \; x \to -\infty, \end{cases} \tag{17.29}$$

  which again corresponds to an $A\mathcal{U}/B\mathcal{U}$ (definition 17.4) type scheduling policy.

17.3 ⊛ Numerical Experiments

In this section the Restless Bandit heuristic will be compared with the optimal policy computed by Ha [22] for the two-items flexible facility problem. In a second group of

experiments, we study the case where the machine can produce more than two different items. The discussion will be based on a comparison between the RBP and other classical heuristic policies which we shall briefly recall. In our simulations, the production rates are all equal (i.e. $\mu_j = \mu,\ j = 1, \ldots, N$), the discounting factor is $\beta = 0.01$, and the net inventory is empty initially.

17.3.1 Review of some Priority Rules for Make-to-Stock Productions

The three heuristics that will be compared are:

Definition 17.4 (The $\mathbf{h_j\mu/b_j\mu}$ Rule). *Let b_j be the cost rate for backorder type j items, h_j the storage cost rate for type j items and μ_j the production rate of type j items. Then, the $h_j\mu/b_j\mu$ reads as follows:*

a) **If demands are backordered:** *Produce the item with the largest $b_j\mu_j$ among all products for which backorder exists.*

b) **If no demands are backordered:** *Produce the item with the smallest $h_j\mu_j$ among all products for witch the inventory levels are under their hedging stock d_j^*.*

Remark: The static $h_j\mu/b_j\mu$ heuristic is fully myopic in the sense that it directly minimizes the instantaneous cost $h(x)$.

Definition 17.5 (The Switching Rule). *The switching rule is obtained by modifying the $h_j\mu/b_j\mu$ rule as follows:*

a) **If demands are backordered:** *Produce the item with the largest $b_j\mu_j$ among all products for which backorder exists (similar to the static $h_j\mu/b_j\mu$ rule).*

b) **If no demand is backordered:** *Produce the item with the largest $b_j\mu_j(1 - x_j/d_j^*)$ value if positive or let the server be idle.*

Remark: The quantity $(1 - x_j/d_j^*)$ can be interpreted as the proportion of unfilled stock. The larger it is, the more probable the backordering of a product. This heuristic implies the existence of a linear switching curve in the positive quadrant of the state space that ends at the hedging point

$$\vec{d^*} = (d_1^*, \ldots, d_N^*).$$

Definition 17.6 (The Priority Index Policy). *The Priority Index Policy is based on the RBP indices given by equation (17.14). For this heuristic, we apply the renormalization procedure of μ given by equation (17.9), with $T = \frac{1}{\mu}$. Unlike the first two policies, this third one is not myopic, as the priority indices take into account the potential cost that can be generated in the future.*

17.3.2 ⊛ Numerical Results

In our experimentation, we measure the total average discounted cost over 4000 simulation runs. The horizon H is chosen to be large enough to guarantee the results to be invariant on H (the presence of the discounting factor makes this possible). For the $h\mu/b\mu$ policy and for the switching rule, we have chosen the hedging stocks which minimize the total discounted cost. To find the hedging stocks for a machine able to produce N types of items, we have used of a N-dimensional search around the optimal hedging

stock, known for a single item problem.

Experiment 1):

The number of item types is: $N = 2$
The demand rates are: $\lambda_1 = 0.4$; $\lambda_2 = 0.5$
The production rate is: $\mu = 1$
The costs are: $B_1 = 30$; $B_2 = 40$; $A_1 = 1$; $A_2 = 1$

With $\beta = 0.01$, the optimal policy has been derived numerically in [22]. It is given by a switching curve and a hedging point $\vec{d}^*$ as follows: When $x < 0$, the optimal policy commands to engage the item having the largest $B_j\mu_j$. When $x \geq 0$, the switching curve is almost equal to the straight line $y = x + 1$ and ends at the hedging levels $\vec{d}^* = (9, 11)$ (i.e. the hedging stock is $d_1^* = 9$ and $d_2^* = 11$).

	Hedging	Cost	Percentage more
Ha	$(9, 11)$	7401	optimal
Restless	(5,8)	7437	$0.5\% \geq$ optimal
Switching	$(4, 7)$	7639	$3.2\% \geq$ optimal
$h\mu/b\mu$	$(1, 8)$	7838	$5.9\% \geq$ optimal

Experiment 2):

The number of item types is: $N = 3$
The demand rates are: $\lambda_j = 0.3$; $j \in \{1, \ldots, 3\}$
The production rate is $\mu = 1$
The cost are $B_1 = 80$; $B_2 = 90$; $B_3 = 100$; $A_1 = 3$; $A_2 = 2$; $A_3 = 3$

	Hedging	Cost	Percentage more
Restless	(4, 4 ,4)	19706	Best
Switching	$(4, 5, 4)$	20763	$5.3\% \geq$ Restless
$h\mu/b\mu$	$(0, 1, 8)$	24672	$25.2\% \geq$ Restless

Experiment 3):

The number of item types is: $N = 4$
The demand rates are: $1/\lambda_1 = 4$; $1/\lambda_2 = 4.1$; $1/\lambda_3 = 4.2$; $1/\lambda_4 = 4.3$
The production rate is $\mu = 1$
The cost are $B_1 = 40$; $B_2 = 30$; $B_3 = 20$; $B_4 = 10$; $A_j = 1$, $j \in \{1, \ldots, 4\}$

	Hedging	Cost	Percentage more
Restless	(2,3,3,3)	7745	Best
Switching	$(2, 3, 3, 4)$	8366	$10.4\% \geq$ Restless
$h\mu/b\mu$	$(1, 1, 3, 3)$	8983	$18.6\% \geq$ Restless

Remark: As noted in [22], we also find that, under the $h\mu/b\mu$ policy, the optimal discounted cost is realized for a strongly uneven distribution of the hedging stock levels (i.e. the capacity of the stock are very different from one type of item to another).

18

Summary

Using the relaxed version of the "Restless Bandit" problem, we are able to compute explicitly the generalized Gittins priority indices for several underlying stochastic processes governing the dynamics of the arms. These explicit expressions are then used in the context of production manufacturing to discuss the dynamic scheduling of jobs in a flexible shop floor. A direct comparison with the optimal policy, known for the two products case, shows that Restless priority indices yield a scheduling policy which is generally close to the optimal one. In particular, if backorder exists, the priority index rule reduces to the scheduling policy which command to produce the item with the largest $b\mu$ (i.e. cost rate for late delivery) among the backordered items. This is consistent with the scheduling rules derived in Wein [48] and Ha [22]. In all simulation experiments performed, we observe that the Restless priority index rule is always better or at least equivalent to the previously studied scheduling heuristics. In any case, the Restless Priority Index Policy performs much better than any purely myopic policy.

Part V

Conclusion of the Thesis and Perspective.

Conclusions

The solution of design and operation modes of complex production systems are often required (by industrialists) in very short time scales. While "rules of thumb" and experience are both essential, the complexity of the issues do nevertheless incite production engineers not to neglect the power of formal and simulation approaches. As a rule, simulation models are largely favoured on the shop-floor level. This is due to their ability to directly match the particularities of the installation under investigation. The very presence of such abundant detailed features often precludes a basic, synthetic and conceptual understanding of the dynamics governing the production/consumption flows. We do however feel confident that significant progresses in the optimization issues to be resolved can be gained via the development of formal models (i.e. mathematical decision models). This approach which is complementary to simulation is the one considered here. Clearly, our work does not provide the shop-floor actors with ready to use answers to many of their questions. Nevertheless it is hoped that for instance the **concept of priority index**, due to its natural and intuitive meaning, will be of interest to decision makers. Hence, such priority indices could be included in the management toolbox in the next future.

Our mathematical modeling belongs to the sequential decision problems known as the Multi-Armed Bandit Problem (MABP). In this general framework the following models are focussed on:

- the classical MABP (without switching cost)
- the Restless Bandit problem,
- the MABP with switching penalties.

In particular, the relevance of these three problems for the Flexible Manufacturing System (FMS) is studied. Our contribution can be summarize as follows:

- The form of the Gittins indices where the evolution of the MABP is given by a piecewise deterministic process which is intrinsically non-Markovian has been computed explicitly. This is among the few classes of non-Markovian examples in the literature for which the Gittins indices can be computed explicitly.

A study of the MABP is of interest because it can be optimally solved by the Priority Index Policy. This simple policy is convincing to continue our work toward a generalization that would yield efficient solutions for more complex decision-making problems. We therefore concentrated on RBP (a generalized form of MABP) for which an efficient Priority Index heuristic called the "Whittle relaxation" already exists:

- RBP with several underlying random dynamics relevant for the production engineering context, (e.g. diffusion processes as well as birth and death processes) have been studied. Explicit generalized priority indices were obtained and the resulting dynamic scheduling was compared with exact results. Finally, using the RBP, we proposed a sub-optimal heuristic solving the multi-items flexible make-to-stock production problem when switching penalties can be neglected.

While the FMS are well implanted almost any large manufacturing unit, the inherent presence of setup costs in FMS to this day still foils the discovery of an optimal scheduling policy. Looking through available literature we rapidly discovered that the contribution to

decision making problems in presence of switching penalties remains mostly unexplored. We therefore decided to focus the end of this thesis on MABP with switching penalties. In particular we concentrated our efforts on the MABP with setup costs and/or time delays:

- Significant progress has been made by constructing a new class for which the optimal policy can be explicitly constructed by recursion. Using this optimal derivation, we then proposed a heuristic, that approaches the optimal policy for general MABP with switching penalties.

Perspectives

The material presented in this thesis suggest two direction axis to be investigated:

a) **RBP with switching penalties:**

In parts II and IV we assumed that the switching penalties can be neglected. Unfortunately, the scheduling problems in presence of switching penalties occur in numerous applications and especially in manufacturing systems. In particular, a complete definition of optimal scheduling for the FMS required to study the RBP with switching penalties. It is then mandatory to construct an efficient heuristic which holds for decision problems in presence of switching penalties. Hence a natural extension of the thesis is to study the RBP with switching costs and/or time delays. In particular, one will try to construct a generalization of the Priority Index Policy for the RBP with switching penalties.

b) **Projects dependent switching penalties:**

In part III we supposed that the switching costs depend neither on the original project nor on the project switched to. When the switching costs depend on both the original project and the project switched to (i.e. $C_{j\to k} \neq C_{u\to v}$ for some $j \neq u$ or $k \neq v$) it would be interesting to study the modification incurred in the optimal policy. We can explore a possible simple generalization of the GIH taking into account this problem as follows:

Define the average cost

$$C = \frac{1}{N(N-1)} \sum_{\substack{j \neq k \\ j,\, k = 1}}^{N} C_{j\to k}.$$

Then, for each project j derive its continuation index $\nu c_j(X_j(t_0))$ defined by equation (12.7) and derive the $N-1$ switching indices $\nu s_{j\to k}(X_j(t_0))$, $k \in \{1,\ldots,N\} \setminus \{j\}$ define as:

$$\nu s_{j\to k}(X_j(t_0)) = \sup_{\pi\in\mathcal{U}} \frac{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi - 1} \beta^{t_i} h_j(X_j(t_i)) - C_{k\to j} - C\beta^{t_{\tau_\pi}}\right\}}{E_{x_j}\left\{\sum_{i=0}^{\tau_\pi - 1} \beta^{t_i}\right\}}. \tag{18.1}$$

A generalization of the GIH would then be:

Definition 18.1 (GIH). *Suppose that the DM is initially engaged on project j then the generalized GIH reads:*

"Engage project j as long as its continuation index is greater or equal than all the switching indices:

$$\nu s_{j\to k}(X_j(t_0)),\ k \in \{1,\ldots,N\} \setminus \{j\}.$$

As soon as the continuation index of project j falls below a switching index $\nu s_{j\to k}(X_j(t_0))$, switch to the project $l \neq j$ having the greatest switching index $\nu s_{j\to l}(X_j(t_0))$ and engage it immediately."

Remerciement

Aujourd'hui, mercredi 16 août 2003, je suis un homme heureux! Je viens de réaliser mon rêve d'enfant, je me suis réveillé avec ce sourire au coin des lèvres et cette pensée dans ma tête: "Je suis docteur..." Cette joie je la dois aux nombreuses personnes qui m'ont soutenu, aidé et avec lesquelles j'ai travaillé. Merci de tout coeur à:

- Max-Olivier Hongler sans qui cette thèse n'aurait jamais vu le jour. Merci pour tout ce que tu as fait pour moi et merci pour tous ces moments privilégiés que nous avons passé ensemble.

- Roger Filliger pour les heures de discussion dont tu m'as fait cadeau. Merci pour ta disponibilité et pour les kilos que je n'ai pas pris...

- Karine Genoud pour ton travaille ton sourire et ta bonne humeur. Sans les personnes de l'ombre il n'y aurait pas de cirque, sans toi il n'y aurait pas de LPM...

- Jacques Jacot pour m'avoir accueilli dans son groupe. Merci pour ta franchise lors de notre premier entretien, pour toutes les discussions que nous avons eues et pour tes précieux exemples industriels.

- Haya Kaspy qui en spécialiste a amélioré substantiellement ma compréhension des problèmes de contrôles optimaux et des difficultés liées à l'introduction des coûts de changement dans les problèmes de décision.

- Robert Dalang pour la lecture attentive de ma thèse et en particulier pour la correction du théorème 5.5.

- Stanley B. Gershwin pour nos discussions enthousiastes et tous tes précieux conseils.

- Stanley B. Gershwin, Alain Haurie, Robert Dalang, Hannes Bleuler, Jacques Jacot et Max-Olivier Hongler qui ont composé mon jury. Merci pour vos stimulantes remarques et votre disponibilité.

- Gerline Brandès, Art Retti, Gavin Seal, Yuri Lopez De Meneses, Julian Randall, Max-Olivier Hongler, Roger Filliger et Bernard Dusonchet pour la relecture de ma thèse.

- Ma famille pour son soutien et sa patience

- Sonia, Nico, Christèle, Isa, Karine, Julian, Véro, Mathieu, Virginie, Yves, Yuri, Max, Roger, Thierry, Carine, Jess, Pascal, Malick, Mitch, France, Greg, Sandra, Jo, François, Gavin, Valérie, Jérôme, Fanny, Cédric, Eric, Alexia, Laurent, Ludivine, Dana, Ion, Aline, Arlette, Did, Clelia, Sophie, Bernard, Marie-Claude, Anne-Marie, Hélène, Jean, Stéphane, Christian, Mario, Ciro, Styve, Fabien, Giancarlo, Svend, David plus tous ceux que j'oublie, pour avoir mis du bonheur dans ma vie.

- Et enfin the last but not least Stefano Retti pour son amitié.

Part VI

Appendices

A

Optimality of the Priority Index Policy for the MABP without switching penalties

Besides the original proof of the optimality of the priority index policy given by Gittins himself in 1974 [19], several other elegant proofs have been written. Due to its simplicity, we will present the proof given by Walrand [46].

Theorem A.1. *The optimal policy for an N-armed Markov MABP in discrete time (without switching penalties) is the "Priority Index Policy" (definition 4.1) with the index value defined by equation (4.3) (i.e. the Gittins index).*

Proof: Without loss of generality, let us make the following assumptions:

- $t_0 = 0$ and the decision times arise each unit of time (i.e. $t_i = i \in \mathbb{N}$),
- initially, the project 1 has the largest index value, i.e.

$$\theta := \nu g_1(X_1(t_0)) \geq \nu g_j(X_j(t_0)),\ 1 < j \leq N,$$

- τ_{π^*} is the optimal stopping time, which achieves the supremum in equation (4.3).

By subtracting θ from each $h_j(x)$, one can assume, without loss of generality, that $\nu g_1(X_1(t_0)) = 0$.

With the above assumption, it will be shown that there is an optimal policy that commands to engage the project 1 at time t_0. Due to the Markov property, we can repeat this argument at any subsequent times, and this will prove the theorem. The basic idea is to show by an interchange argument that an arbitrary policy can be improved by a modification such that project 1 is engaged at time t_0.

Let $\pi \in \mathcal{U}$ be an arbitrary policy. Following π, we engage at each decision time t_i one of the N projects and get its reward. We can therefore define the policy as its sequence of engagements or similarly as its sequence of rewards. Let us focus on the decision times at which the policy π engages the project 1. Write the sequence of rewards of π as:

$$\pi \ : \ \Big([\ldots]_0, h_1(X_1(t_0^1)), [\ldots]_1, h_1(X_1(t_1^1)), [\ldots]_2, h_1(X_1(t_2^1)), \ldots$$

$$\ldots, h_1(X_1(t^1_{\tau_{\pi^*}-1})), h_k(X_k(t_l^k)), [\ldots]_l\Big),$$

where $[\ldots]_n$ indicates the rewards obtained when project 1 is not engaged (note that some $[\ldots]_n$ may be equal to 0), $h_1(X_1(t_n^1))$ denotes the reward received when engaging the project 1 for the $(n+1)$-th time and t_n^j is the time at which project j is engaged for

the $(n+1)$-th time under policy π.

For ease of notation let us write $T := t^1_{\tau_{\pi^*}-1}$ (i.e. the time at which the reward $h_1(X_1(t^1_{\tau_{\pi^*}-1}))$ is received). Consider now the policy $\tilde{\pi}$ which is a modification of policy π such that it engages at time t_0 the project 1 during a time τ_{π^*}, then engages until time T the other projects in the same order as policy π and coincides for $t \geq T$ with policy π. The sequence of reward of $\tilde{\pi}$ is as follows:

$$\tilde{\pi} \ : \ \Big(h_1(X_1(t^1_0)), h_1(X_1(t^1_1)), \ldots, h_1(X_1(t^1_{\tau_{\pi^*}-1})), [\ldots]_0, [\ldots]_1, \ldots, h_k(X_k(t^k_l)), [\ldots]_l\Big).$$

By construction, the sequences of rewards of policy π and $\tilde{\pi}$ coincide after time T. Denote by R, respectively $\tilde{R}$, the expected sequences of rewards up to time T given by policy π respectively $\tilde{\pi}$. To prove the Theorem it suffice to show that $\tilde{R} \geq R$. Let us start by the following definition:

- t^j_n is the time when project $1 \leq j \leq N$ is selected for the $(n+1)$-th time by policy π.
- $\tilde{t}^j_n$ is the time when project $1 \leq j \leq N$ is selected for the $(n+1)$-th time by policy $\tilde{\pi}$.
- $m_j + 1$ is the number of times that project j is selected up to time T by the policies π and $\tilde{\pi}$.

Observe that the value $h_j(X_j(t^j_n))$ under policy π is by definition equivalent to the value $h_j(X_j(\tilde{t}^j_n))$ under policy $\tilde{\pi}$. Now,

$$\begin{aligned}\tilde{R} - R = E\left\{\sum_{i=0}^{\tau_{\pi^*}-1} \beta^{\tilde{t}^1_i} h_1(X_1(\tilde{t}^1_i))\right\} - E\left\{\sum_{i=0}^{\tau_{\pi^*}-1} \beta^{t^1_i} h_1(X_1(t^1_i))\right\} + \\ + \sum_{j=2}^{N} E\left\{\sum_{i=0}^{m_j} (\beta^{\tilde{t}^j_i} - \beta^{t^j_i}) h_j(X_j(t^j_i))\right\}.\end{aligned} \tag{A.1}$$

Remark:

- As $\tilde{t}^1_i = i$ and $\nu g_1(X(t_0)) = 0$, the first term on the right-hand side of equation (A.1) is equal to zero (it corresponds to the numerator of equation (4.3)).
- Since $\nu g_1(X(t_0)) = 0$, the second term is nonpositive. Indeed, it corresponds to the expected reward obtained by engaging project 1 with inserted idle time. Since these idle times cannot improve the maximum expected reward which is nonpositive, it follows that this term is nonpositive.

It only remains to show that the last term corresponding to the reward received by each $j = 2, 3, \ldots, N$, is nonnegative. Define σ^j_i as:

$$\beta^{\sigma^j_i} = \beta^{t^j_i} - \beta^{\tilde{t}^j_i}.$$

Then the σ^j_i are random times such that $\sigma^j_{i+1} \geq \sigma^j_i + 1$. To verify this fact, notice that

$$\beta^{\sigma^j_i - i} = \beta^{t^j_i - i}(1 - \beta^{\tilde{t}^j_i - t^j_i}).$$

By the construction of the policy $\tilde{\pi}$, it follows directly that $t^j_i - i$ and $t^j_i - \tilde{t}^j_i$ are non-decreasing in i for $j \neq 1$. Hence $\beta^{\sigma^j_i - i}$ is non-increasing in i which implies that $\sigma^j_i - i$ must be non-decreasing in i. Moreover the value σ^j_i is a function of the project $X_j(\cdot)$ up

to time t_i^j and of the evolution of the other projects. It is therefore a stopping time of project j by the independence of the project.

We can then rewrite the the last term of equation (A.1) as:

$$E\left\{\sum_{i=0}^{m_j}(\beta^{t_i^j}-\beta^{\bar{t}_i^j})h_j(X_j(t_i^j))\right\}=E\left\{\sum_{i=0}^{m_j}\beta^{\sigma_i^j}h_j(X_j(t_i^j))\right\} \tag{A.2}$$

One may view equation (A.2) as being the expected reward received when engaging project j with inserted idle times, the idle times being randomized by the evolution of the other projects. Again, as these idle times cannot improve the maximum expected reward which is nonpositive, equation (A.2) is nonpositive.

□

B

⊛ Position of the Hedging Stock

Here we compute the optimal position of the hedging level for a single class make-to-stock server with Markov dynamics $X(t)$.

In section 17.2.1, we saw that the optimal policy for a single item make-to-stock problem is a hedging stock policy π^{d^*} with a hedging lever d^*. This policy is defined as follows: "*Produce when the stock level is below the level d^* or let the machine be idle*". Under such a policy, the stochastic process:

$$Y(t) = d^* - X(t)$$

is isomorphic to a $M/M/1$ queue [35]. Let us denote by $J^{d^*}(x)$, the cost incurred under policy π^{d^*}. Optimality of the hedging stock implies that the discrete derivative with respect to d^* vanishes, namely:

$$J^{d^*}(x) - J^{d^*+1}(x) = 0. \tag{B.1}$$

To solve equation (B.1) let us first recall that:

$$J^{d^*}(x) = E_x \int_0^\infty e^{-\beta t} h(X(t)) dt.$$

Let τ be an exponentially distributed random variable with mean $1/\beta$ independent of $X(t)$ and $h(x)$. Then

$$J^{d^*}(x) = \frac{1}{\beta} E_x E_\tau \left[h(X(\tau)) \right]. \tag{B.2}$$

Permutating the expectation operators in equation (B.2), we have:

$$\beta J^{d^*}(x) = E_\tau \left[\sum_{i=-\infty}^{d^*} h(i) P(X(\tau) = i \,|\, X(0) = x) \right]. \tag{B.3}$$

Letting $Y(\tau) = d^* - X(\tau)$ we obtain:

$$\beta J^{d^*}(x) = E_\tau \left[A \sum_{y=0}^{d^*-1} (d^* - y) P(Y(\tau) = y \,|\, Y(0) = 0) + B \sum_{y=d^*+1}^{\infty} (y - d^*) P(Y(\tau) = y \,|\, Y(0) = 0) \right].$$

From equation (B.1) we obtain:

$$0 = \beta J^{d^*+1}(x) - \beta J^{d^*}(x) = E_\tau \left[(A+B) \left(1 - \sum_{y=d^*+1}^{\infty} P(Y(\tau) = y \,|\, Y(0) = 0) \right) - B \right].$$

Computing the expectation with respect to τ, we find that:

$$0 = \beta J^{d^*+1}(x) - \beta J^{d^*}(x) =$$

$$A - (A+B)\beta \int_0^\infty e^{-\beta t} \sum_{y=d^*+1}^{\infty} P\left(Y(t) = y \,|\, Y(0) = d^* - x\right) dt.$$

It is known that when $x = d^*$ (i.e. $Y(0) = 0$), the Laplace transform $f(\beta, y)$ of the transient probability density of the $M/M/1$ queue $P\left(Y(t) = y \,|\, Y(0) = d^* - x\right)$ reads simply as (see [33] for example):

$$f(\beta, y) = \frac{(1-\omega_-)\omega_-^y}{\beta},$$

where:

$$\omega_- = \frac{(\beta + \lambda + \mu) - \sqrt{(\beta+\lambda+\mu)^2 - 4\lambda\mu}}{2\mu}.$$

So we obtain:

$$\beta \int_0^\infty e^{-\beta t} \sum_{y=d^*+1}^{\infty} P\left(Y(t) = y \,|\, Y(0) = d^* - x\right) dt = \omega_-^{d^*+1},$$

and we end with:

$$0 = A - (A+B)\omega_-^{d^*+1} \quad \Rightarrow \quad d^* = \left\lfloor \frac{1}{\ln(\omega_-)} \ln\left(\frac{A}{A+B}\right) \right\rfloor.$$

C

⊛ Optimal Cost Functions - Markov Chain Dynamics

Here we derive the optimal cost functions $J_a(x,\gamma)$, $J_p(x,\gamma)$ for a single class make-to-stock server where the dynamics is given by a Markov chain in continuous time.

We saw in section 8.1.3 that the optimal cost functions for the γ-penalized problem obeys to equation (8.16):

$$\begin{cases} \beta J_a(x,\gamma) = h(x) + \lambda J_a(x-1,\gamma) + \mu J_a(x+1,\gamma) - (\lambda+\mu) J_a(x,\gamma) \\ \\ \beta J_p(x,\gamma) = h(x) + \lambda J_p(x-1,\gamma) - \lambda J_p(x,\gamma) + \gamma. \end{cases}$$

This system is linear and the general solutions of the homogenous system are

$$\begin{aligned} J_a(x,\gamma) &= C_{a^+}(w_+)^x + C_{a^-}(w_-)^x, \\ J_p(x,\gamma) &= C_p(w_0)^x, \end{aligned}$$

where C_{a^+}, C_{a^-}, C_p are integration constants and

$$\begin{aligned} w_+ &= \tfrac{(\beta+\lambda+\mu)+\sqrt{(\beta+\lambda+\mu)^2-4\lambda\mu}}{2\mu}, \\ w_- &= \tfrac{(\beta+\lambda+\mu)-\sqrt{(\beta+\lambda+\mu)^2-4\lambda\mu}}{2\mu}, \\ w_0 &= \tfrac{\lambda}{\lambda+\beta}. \end{aligned} \tag{C.1}$$

The particular solutions correspond to engage the server forever or to let the server be idle forever. For the active case we get:

$$\begin{aligned} &S_a(x,\gamma) = E_x\left[\textstyle\int_0^\infty e^{-\beta s} h(X(s))ds\right] = \\ &\textstyle\int_0^\infty e^{-\beta s} \sum_{k=-\infty}^{\infty} h(k) P\{X(s)=k \,|\, X(0)=x\} ds = \\ &\textstyle\int_0^\infty e^{-\beta s} \sum_{k=-\infty}^{\infty} h(k) P\{X(s)=k-x \,|\, X(0)=0\} ds. \end{aligned} \tag{C.2}$$

To compute the transition probability density $P\{X(s)=k-x \,|\, X(0)=0\}$, consider a space-homogeneous Markov chain process $X(t)$ with parameter λ and μ. Define $\rho = \frac{\lambda}{\mu}$ and $P_n(t) = P\{X(t)=n \,|\, X(0)=0\}$. We know that $P_n(t)$ follows the equation (see for example [41]):

$$\frac{d}{dt}P_n(t) = -(\lambda+\mu)P_n(t) + \mu P_{n+1}(t) + \lambda P_{n-1}(t). \tag{C.3}$$

Define:

$$P_n(t) = Q_n(t)\rho^{-\frac{n}{2}} e^{-(\lambda+\mu-2\sqrt{\lambda\mu})t}$$

in terms of which the equation (C.3) becomes:

$$\frac{d}{dt}Q_n(t) = \sqrt{\lambda\mu}(Q_{n+1}(t) + Q_{n-1}(t) - 2Q_n(t)).$$

The solution of this last equation reads (see for example [16]):

$$Q_n(t) = e^{-2\sqrt{\lambda\mu}t}\mathbb{I}_{|n|}(2\sqrt{\lambda\mu}t)$$

with $\mathbb{I}_n(x)$ being the modified Bessel function:

$$e^{\frac{1}{2}x(t+1/s)} = \sum_{k=-\infty}^{\infty} t^k \mathbb{I}_n(x).$$

Hence, we obtain:

$$P_n(t) = \rho^{-\frac{n}{2}} e^{-(\lambda+\mu)t}\mathbb{I}_{|n|}(2\sqrt{\lambda\mu}t)$$

Using the Laplace transform of $P_n(t)$, which enters directly into equation (C.2), we end with:

$$S_a(x,\gamma) = \sum_{k=-\infty}^{\infty} \frac{h(k)\rho^{-\frac{k-x}{2}}}{\sqrt{(\beta+\lambda+\mu)^2 - 4\lambda\mu}} \left[\frac{(\beta+\lambda+\mu) - \sqrt{(\beta+\lambda+\mu)^2 - 4\lambda\mu}}{2\sqrt{\lambda\mu}}\right]^{|k-x|}$$

From equation (C.1) and the fact that $w_+ w_- = \frac{\lambda}{\mu} = \rho$, we can show that $S_a(x,\gamma)$ takes the form:

$$S_a(x,\gamma) = \frac{1}{\mu(w_+ - w_-)}\left\{h(x) + (w_-)^x \sum_{k=-\infty}^{x-1} h(k)(w_-)^{-k} + (w_+)^x \sum_{k=x+1}^{\infty} h(k)(w_+)^{-k}\right\}. \tag{C.4}$$

Along the same lines, when the server is idle forever, we obtain:

$$S_p(x,\gamma) = \int_0^{\infty} e^{-\beta s} \sum_{k=-\infty}^{\infty} (h(k)+\gamma)(P\{X(s) = k - x \,|\, X(0) = 0\}ds.$$

In this case $P\{X(s) = k - x \,|\, X(0) = 0\}$ is a Poisson process and we end with:

$$S_p(x) = (w_0)^{x+1} \sum_{k=-\infty}^{x} \frac{(h(k)+\gamma)(w_0)^{-k}}{\lambda}. \tag{C.5}$$

D

⊛ Index Obtained for a Piecewise Linear Running Cost

Here we compute the index for the single class make-to-stock server problem described in section 17.2.1 where the dynamics is a continuous time Markov chain and the cost rate function $h(x)$ is piecewise linear:

$$h(x) = \begin{cases} +Ax \;\; ; \;\; A > 0 \quad \text{if } x \geq 0 \\ -Bx \;\; ; \;\; B > 0 \quad \text{if } x < 0. \end{cases}$$

We have shown in section 17.2.1 that the cost functions $J_a(x)$ and $J_p(x)$ are:

$$J_a(x,\gamma) = C_a(w_+)^x + S_a(x,\gamma)$$
$$J_p(x,\gamma) = C_p(w_0)^x + S_p(x,\gamma),$$

where $S_a(x)$ and $S_p(x)$ are given by equation (C.4) and equation (C.5) respectively.

Using the shape of $h(x)$, we derive the closed form of $J_a(x,\gamma)$ and $J_p(x,\gamma)$. For the region $x \geq 0$, the summation formula for geometric series implies:

$$J_a(x,\gamma) = \tfrac{1}{\beta^2\mu(w_+ - w_-)}\Big\{2^{-x-1}\Big[(A+B)\Big((\lambda-\mu)^2 + \beta(\lambda+\mu)\Big)(2w_-)^x +$$

$$+\mu(w_+ - w_-)\Big(2^{x+1}A(\beta x - \lambda + \mu) + A(\lambda-\mu)(2w_-)^x +$$

$$+B(\lambda-\mu)(2w_-)^x + 2C_{a^+}\beta^2(2w_+)^x\Big)\Big]\Big\},$$

and

$$J_p(x,\gamma) = \tfrac{(\beta+\lambda)(A+\beta-A\lambda+\beta\gamma)+(w_0)^x\left((\beta+\lambda)(C_p\beta^2+(A+B)\lambda)+\beta^2\gamma\right)}{\beta^2(\beta+\lambda)}.$$

Similarly, for the region $x < 0$ region, we obtain:

$$J_a(x,\gamma) = \tfrac{1}{\beta^2\mu(w_+ - w_-)}\Big\{2^{-x-1}\Big[(A+B)\Big((\lambda-\mu)^2 + \beta(\lambda+\mu)\Big)(2w_+)^x +$$

$$+\mu(w_+ - w_-)\Big(-2^{x+1}B(\beta x - \lambda + \mu) + \big(2C_{a^+}\beta^2 +$$

$$-(A+B)(\lambda-\mu)\big)(2w_+)^x\Big)\Big]\Big\},$$

and

$$J_p(x,\gamma) = \tfrac{B(-\beta x+\lambda)+\beta(C_p\beta(W_0)^x+\gamma)}{\beta^2}.$$

Using the extended smooth-fit principle given in equation (8.8) and the definition given in equation (17.13), we can derive the index $\nu(x)$ in the form given by equation (17.14).

E

⊛ Proof of Proposition 14.3

Proposition 14.3: For any initial condition, the optimal policy, for a two-armed DMABP in class $\mathcal{Z}$, commands to switch only a finite number of times.

Proof:
In the following, we use the notation $[\,\cdot \mid X_j(t) = x_j]$ to indicate that the project j is in state x_j at time t.

The space of initial conditions $(x_1, x_2, j) \in \mathbb{R}^2 \times \{1, 2\}$ can be split into two disjoint subsets:

a) The set of initial conditions $(x_1, x_2, j) \in \Lambda$, defined by one of the two following properties:

   i) The DM is initially engaged on arm $j = 1$ and:

$$\left[\int_0^\infty e^{-\beta t} h_2(X_2(t))\, dt \;\middle|\; X_2(0) = x_2\right] - C \leq \frac{\Gamma_1}{\beta}$$

   i.e. the reward gained by initially paying the switching cost and then by engaging project 2 forever is smaller than the smallest possible reward gained by engaging project 1 alone forever.

   ii) The DM is initially engaged on arm $j = 2$ and:

$$\left[\int_0^\infty e^{-\beta t} h_1(X_1(t))\, dt \;\middle|\; X_1(0) = x_1\right] - C \leq \frac{\Gamma_2}{\beta}$$

   i.e. the reward gained by initially paying the switching cost and then by engaging project 1 forever is smaller than the smallest possible reward gained by engaging project 2 alone forever.

b) The complementary set $\Lambda' = \left\{\mathbb{R}^2 \times \{1, 2\}\right\} \setminus \Lambda$.

Note that, for any initial condition $(x_1, x_2, j) \in \Lambda$, the optimal policy clearly commands, by definition, to engage project j forever. Moreover, starting with an initial condition in Λ, every subsequent state reached when following the optimal policy, also belongs to Λ. Indeed, for DMABP we have:

$$h_j(X_j(t_2)) \leq h_j(X_j(t_1)),\ \forall t_2 \geq t_1,\ j = 1, 2.$$

E ⊛ Proof of Proposition 14.3

For any initial conditions $(x_1, x_2, j) \in \Lambda$, the optimal policy commands to engage project j forever. Hence, it satisfies the assertion of proposition 1.

It then remains to prove proposition 1 for initial conditions in Λ'. To do this, let us define $\{t_i^1\}$, $i = 1, 2, \ldots$, the sequence of times at which the optimal policy gives the order to switch from project 1 to 2 and $\{t_i^2\}$, $i = 1, 2, \ldots$, the sequence of times at which the optimal policy gives the order to switch from project 2 to 1 (see the Fig.E.1).

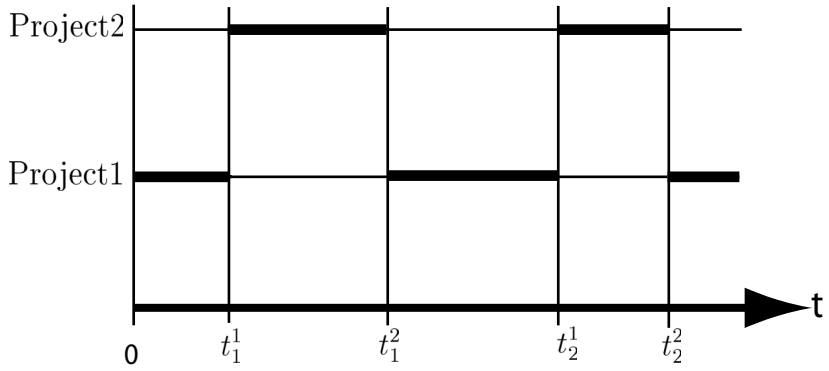

Fig. E.1. The sequence of switching times t_i^1 and t_i^2 with the DM initially engage on project 1.

Let j be the project initially engaged and $\bar{j}$ the disengaged project. Define $\mathcal{H}j$ and $\mathcal{H}\bar{j}$ to be the cumulated sojourn times spent on project j, respectively on project $\bar{j}$, under the optimal policy. For an infinite horizon, we have that $\mathcal{H}j + \mathcal{H}\bar{j} = \infty$ and one has necessarily one of the three following alternatives:

i) $\mathcal{H}1 = \infty$ and $\mathcal{H}2 = \infty$,
ii) $\mathcal{H}1 < \infty$ and $\mathcal{H}2 = \infty$,
iii) $\mathcal{H}1 = \infty$ and $\mathcal{H}2 < \infty$.

Let us assume, without loss of generality, that $\Gamma_{\bar{j}} \geq \Gamma_j$ and assume first that:

• Alternative i) holds:

We will show that there exists a time instant $T < \infty$ at which the system enters into the set Λ. This contradicts with the fact that the optimal policy fulfill alternative i) as after time T the optimal policy does not command to switch anymore.

Engaging j alone during [0,T] write $x_j^T = [X_j(T) \mid X_1(0) = x_0]$ for the reached position at time time T. We now show that: $\forall \xi > 0$, $\exists\ T < \infty$ such that the following properties are simultaneously fulfilled:

a) $\left[\int_0^\infty e^{-\beta t} h_j(X_j(t))\, dt \;\middle|\; X_i(0) = x_j^T\right] < \frac{\Gamma_j}{\beta} + \xi$,

b) At time T the DM is engaged on project $\bar{j}$.

By hypothesis, the cumulated sojourn times spent on projects 1 and 2 are infinite. Using equation (14.1), we have:

$$\forall\ \xi' > 0,\ \ \exists\ \tilde{T} > 0,\ \tilde{T} < \infty \text{ such that } \ h_j(X_j(\tilde{T})) < \Gamma_j + \xi'. \tag{E.1}$$

Therefore property a) is satisfied with $\xi = \frac{\xi'}{\beta}$. Moreover, as we consider DMABP, $\forall\ T \geq \tilde{T}$ property a) is necessary satisfied.

Observe that due to the discount factor $e^{-\beta t}$ in equation (3.4), the alternative i) necessarily implies an infinite number of switchings. Note further that any policy that commands to switch an infinite number of times during a finite time interval, incurs an infinite cost and hence cannot possibly be optimal. This implies that an optimal policy only allows a finite number of switches on a finite time horizon. Therefore, when alternative i) is satisfied, we must have that:

$$\forall\ \tilde{T} > 0,\ \exists\ T \geq \tilde{T}$$

such that the optimal policy commands to switch from project j to $\bar{j}$ at time T and hence b) is satisfied.

As a) and b) hold simultaneously, we have:

$$\left[\int_0^\infty e^{-\beta t} h_j(X_j(t))\, dt \;\middle|\; X_i(0) = x_j^T\right] < \frac{\Gamma_j}{\beta} + \xi \leq \frac{\Gamma_{\bar{j}}}{\beta} + \xi.$$

Choosing $\xi < C$, we then have:

$$\left[\int_0^\infty e^{-\beta t} h_j(X_j(t))\, dt \;\middle|\; X_i(0) = x_j^T\right] - C < \frac{\Gamma_{\bar{j}}}{\beta},$$

which implies that $(X_{\bar{j}}(T), X_j(T), \bar{j}) \in \Lambda$ by definition of Λ. Hence, after the time $T < \infty$, the optimal policy never commands to switch anymore and hence $T_j < \infty$ which contradicts the hypothesis.

Let us now prove proposition 1 when:

• **Alternative** ii) **holds:**

Assume, ad absurdum, that the optimal policy π^* commands to switch an infinite number of times. As alternative ii) to hold, we must have:

$$\sum_{i=1}^{\infty} t_i^1 - t_{i-1}^2 = \mathcal{H}1 < \infty.$$

This implies that:

$$\forall \xi > 0,\ \exists\ i\ \ \text{s.t}\ .\ t_i^1 - t_{i-1}^2 \leq \xi.$$

For ξ small enough, we will have:

$$\int_{t_{i-1}^2}^{t_i^1} e^{-\beta t} h_1(X_1(t))\, dt < C.$$

Therefore, the reward gained during the time interval $[t_{i-1}^2, t_i^1[$ is smaller that the switching cost and the policy π^* cannot possibly be optimal. Hence, a contradiction.

• **Alternative** iii) **holds:** Use the same arguments than for alternative ii).

□ Proposition 1

F

⊛ Proof of Proposition 14.4

Proposition 14.4: The optimal policy for a two-armed DMABP in the class $\mathcal{Z}$ is characterized by two switching curves $\mathcal{SO}_{1\to 2}$ and $\mathcal{SO}_{2\to 1}$ defined respectively by two functions, $\tilde{y}: x_1 \mapsto \tilde{y}(x_1)$ and $\tilde{x}: x_2 \mapsto \tilde{x}(x_2)$ (see Figure F.1).

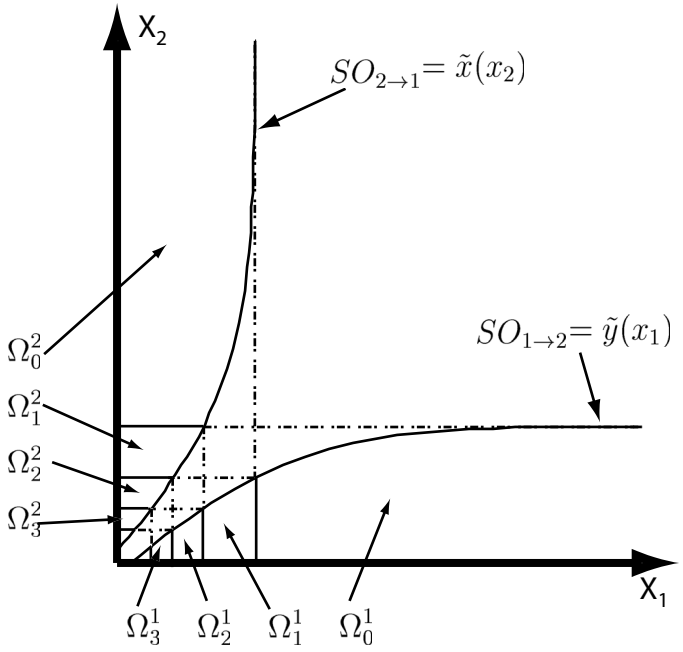

Fig. F.1. The sets Ω_n^j, $j = 1, 2$.

Proof:
Let us start by introducing a few definitions:

- $\Omega_n^1 = \left\{ (x_1, x_2, 1) \in \mathbb{R}^2 \times \{1,2\} \;\middle|\; \text{the optimal policy commands to switch immediately from project 1 to project 2 and then commands to switch exactly n times} \right\}$, $n = 0, 1, 2, \ldots$ (see Fig.F.1).
- $\Omega_n^2 = \left\{ (x_1, x_2, 2) \in \mathbb{R}^2 \times \{1,2\} \;\middle|\; \text{the optimal policy commands to switch immediately from project 2 to project 1 and then commands to switch exactly n times} \right\}$, $n = 0, 1, 2, \ldots$ (see Fig.F.1).

- $V_0(x_1, x_2, j)$ denotes the reward gained by engaging project j forever when starting at position (x_1, x_2, j), namely:

$$V_0(x_1, x_2, j) = \left[\int_0^\infty e^{-\beta t} h_j(X_j(t))\, dt \;\middle|\; X_j(0) = x_j \right].$$

- $V_n(x_1, x_2, j)$ denotes the optimal reward gained when starting at initial position (x_1, x_2, j) and with the assumption that $(x_1, x_2, \bar{j})$ belongs to $\Omega_n^{\bar{j}}$. This is equivalent to say that starting at (x_1, x_2, j), the optimal policy commands to switch exactly n times).

- $T(x_1, x_2, j)$ denotes the time of the first switch from project j to project $\bar{j}$ under the optimal policy and with the initial conditions (x_1, x_2, j).

- $(\bar{x}_1, \bar{x}_2)$ denotes the position reached at the first switch starting at (x_1, x_2, j) and following the optimal policy (see Fig.F.3 for $V_1(x_1, x_2, 2)$).

The dynamic programing principle implies:

$$V_n(x_1, x_2, j) =$$
$$\underbrace{\left[\int_0^{T(x_1,x_2,j)} e^{-\beta t} h_j(X_j(t))\, dt \;\middle|\; X_j(0) = x_j \right]}_{a)} + \underbrace{e^{-\beta T(x_1,x_2,j)} \Big(-C + V_{n-1}(\bar{x}_1, \bar{x}_2, \bar{j}) \Big)}_{b)}. \tag{F.1}$$

The term $a)$ describes the reward received when engaging the project j until time $T(x_1, x_2, j)$. The term $b)$ describes the global reward from time $T(x_1, x_2, j)$ onward (i.e. with initial condition $(\bar{x}_1, \bar{x}_2, \bar{j})$) minus the first switching cost.

Assume, without loss of generality, that the DM is initially engaged on project $j = 1$. Then, by definition:

> $(x_1, x_2, 1) \in \Omega_n^1 \;\Leftrightarrow\; \forall \tau \in \mathbb{R}_+$, *it more rewarding to immediately engage project 2 and then to proceed optimally, than to stay on project 1 during a time* τ*, then to switch to project 2 and finally to proceed optimally.*

Formally, this reads as:

$$\underbrace{-C + V_n(x_1, x_2, 2)}_{a)} \geq \underbrace{\left[\int_0^\tau e^{-\beta t} h_1(X_1(t))\, dt \;\middle|\; X_1(0) = x_1 \right]}_{b)} + \underbrace{e^{-\beta \tau} \Big(-C + V_k(x_1^\tau, x_2, 2) \Big)}_{c)}, \tag{F.2}$$

for some $k \geq 0$ and x_1^τ is the position attained by project 1 when engaged during a time τ. The term $a)$ describes the reward received when immediately engaging project 2 paying the switching penalties and then proceeding optimally. The term $b)$ describes the reward received when engaging the project 1 until time τ. The term $c)$ describes the global reward from time τ) onward (i.e. with initial condition $(x_1^\tau, x_2, 2)$) minus the switching cost.

Lemma F.1. *If at position* $(x_1, x_2, 2)$ *it is optimal to engage project 2 forever, then it is also optimal to engage project 2 forever at every position* $(x_1', x_2, 2)$ *with* $x_1' \geq x_1$.

Proof: Note that $V_0(x_1, x_2, 2) = V_0(x_1', x_2, 2)\ \forall x_1' \geq x_1$. Assume that, starting at position $(x_1', x_2, 2)$, the optimal policy $\tilde{\pi}$ commands to switch at least once. Write $\tilde{V}_0(x_1', x_2, 2)$ for the gain received under policy $\tilde{\pi}$. We then have that $\tilde{V}_0(x_1', x_2, 2) > V_0(x_1', x_2, 2)$. Write $\tilde{\tau}_i$, $i = 1, 2, \ldots$ for the time instants at which the policy $\tilde{\pi}$ command to switch. Starting at $(x_1, x_2, 2)$ denote by $\tilde{V}_0(x_1, x_2, 2)$ the global reward obtained when engaging the projects X_1 and X_2 as policy $\tilde{\pi}$ (i.e. switching the project at time $\tilde{\tau}_i$ see Figure F.2). As $h_j(x)$ is non-increasing,

$$\tilde{V}_0(x_1, x_2, 2) \geq \tilde{V}_0(x_1', x_2, 2) > V_0(x_1', x_2, 2) = V_0(x_1, x_2, 2)$$

which contradicts the optimality of $V_0(x_1, x_2, 2)$.

□

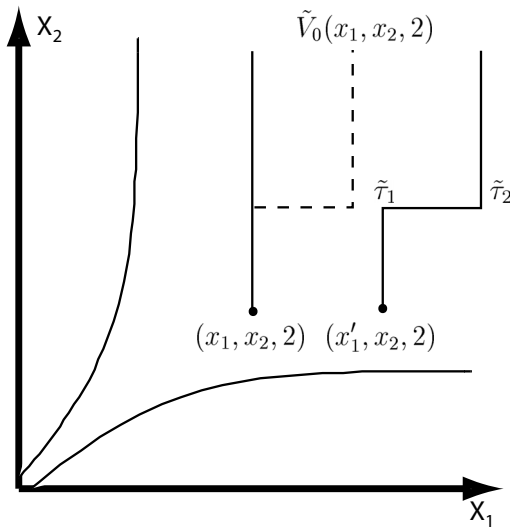

Fig. F.2. Reward $\tilde{V}_0(x_1, x_2, 2)$.

We continue the proof by iteration on n. We will show that given x_1, the position in Ω_n^1 for a given n are those for which $(x_1, z, 1)$ with $z \in]-\infty, \tilde{y}(x_1)]$.

Lemma F.2 (Iteration n=0, for project j=1). *Assume that* $(x_1, x_2, 1) \in \Omega_0^1$, *then it exists* $\tilde{y}(x_1)$ *such that:*

$$\forall z \in]-\infty, \tilde{y}(x_1)],\ (x_1, z, 1) \in \Omega_0^1.$$

Moreover, $\forall (x_1, z', 1)$ *with* $z' > \tilde{y}(x_1)$, *we have* $(x_1, z', 1) \notin \Omega_0^1$.

Proof:
By hypothesis $(x_1, x_2, 1) \in \Omega_0^1$, then using equation (F.2) with $n = 0$ and the Lemma F.1, we have that $\forall \tau \in \mathbb{R}^+$:

$$-C + V_0(x_1, x_2, 2) \geq \left[\int_0^\tau e^{-\beta t} h_1(X_1(t))\, dt \;\middle|\; X_1(0) = x_1 \right] + e^{-\beta\tau}\Big(-C + V_0(x_1^\tau, x_2, 2)\Big). \tag{F.3}$$

As $V_0(x_1, x_2, 2) = V_0(x_1^\tau, x_2, 2)$, equation (F.3) is equivalent to:

$$\frac{\left[\int_0^\tau e^{-\beta t}(h_1(X_1(t)) + \beta C)\, dt \;\middle|\; X_1(0) = x_1\right]}{\int_0^\tau e^{-\beta t}} \leq \beta V_0(x_1, x_2, 2),\ \forall \tau \in \mathbb{R}^+. \tag{F.4}$$

F ⊛ Proof of Proposition 14.4

Dealing with DMABP, Lemma 5.3 implies that $h_j(X_j(t))$ is decreasing in time. Therefore $\forall x_1 \in \mathbb{R}^+$ fixed, the function $V_0(x_1, x_2, 2)$ is decreasing in x_2. Hence, if equation (F.4) holds for $(x_1, x_2, 2)$ (i.e. $(x_1, x_2, 2) \in \Omega_0^1$), it also holds for every points $(x_1, x_2', 2)$ with $x_2' < x_2$. Given x_1, write $\tilde{y}(x_1)$ for the value of x_2 when (F.4) is an equality at position $(x_1, \tilde{y}(x_1), 1)$. It is therefore possible to define the function:

$$\tilde{y}(x_1) \ : \ x_1 \mapsto \tilde{y}(x_1).$$

□ Lemma F.2

Lemma F.3 (Iteration n=0, for project j=2). *Assume that* $(x_1, x_2, 2) \in \Omega_0^2$, *then,* $\exists\ \tilde{x}(x_2)$ *such that:*

$$\forall z \in]-\infty, \tilde{x}(x_2)],\ (x_1, z, 2) \in \Omega_0^2.$$

Moreover, $\forall (x_1, z', 2)$ *with* $z' > \tilde{x}(x_2)$, *we have* $(x_1, z', 2) \notin \Omega_0^2$.

Proof:
Follow the same arguments as in Lemma F.2 .

□ Lemma F.3

Lemma F.4 (Iteration n=1, for project j=1). *Assume that* $(x_1, x_2, 1) \in \Omega_1^1$, *then,* $\exists\ \tilde{y}(x_1)$ *such that:*

$$\forall z \in]-\infty, \tilde{y}(x_1)],\ (x_1, z, 1) \in \Omega_1^1.$$

Moreover, $\forall (x_1, z', 1)$ *with* $z' > \tilde{y}(x_1)$, *we have* $(x_1, z', 1) \notin \Omega_1^1$.

Proof:
By hypothesis $(x_1, x_2, 1) \in \Omega_1^1$, and using equation (F.2) with $n = 1$, we have that $\forall \tau \in \mathbb{R}^+$:

$$-C + V_1(x_1, x_2, 2) \geq \left[\int_0^\tau e^{-\beta t} h_1(X_1(t))\, dt \ \middle|\ X_1(0) = x_1 \right] + e^{-\beta\tau}\Big(-C + V_k(x_1^\tau, x_2, 2)\Big), \tag{F.5}$$

where x_1^τ is the position taken by project 1 after engaging it during a time τ (see Fig.F.3). In view of Fig.F.3, the optimal reward $V_k(x_1^\tau, x_2, 2)$ and $V_0(\bar{x}_1, \bar{x}_2, 1)$ can be decomposed as follows:

$$V_k(x_1^\tau, x_2, 2) = \underbrace{\int_0^{T(x_1, x_2, 2)} e^{-\beta t} h_2(X_2(t))\, dt}_{a)} + e^{-\beta T(x_1, x_2, 2)} V_k(x_1^\tau, \bar{x}_2, 2), \tag{F.6}$$

$$V_0(\bar{x}_1, \bar{x}_2, 1) = \underbrace{\int_0^{\tau} e^{-\beta t} h_1(X_1(t))\, dt}_{b)} + e^{-\beta\tau} V_0(x_1^\tau, \bar{x}_2, 1) \tag{F.7}$$

The term a) describes the reward received when engaging the project 2 from position $(x_1^\tau, x_2, 2)$ until state $(x_1^\tau, \bar{x}_2, 2)$ which is the intersection of both trajectories. The term b) describes the reward received when engaging the project 1 from state $(\bar{x}_1, \bar{x}_2, 1)$ until state $(x_1^\tau, \bar{x}_2, 1)$ which again is the position at which both trajectory meet.
Using equation (F.1) for $V_1(x_1, x_2, 2)$, and using equations (F.6) and (F.7), after calcu-

lations the equation (F.5) can be rewritten as:

$$-C+\left[\int_0^{T(x_1,x_2,2)} e^{-\beta t}h_2(X_2(t))\,dt \;\middle|\; X_2(0)=x_2\right](1-e^{-\beta\tau})+$$
$$-\left[\int_0^{\tau} e^{-\beta t}h_1(X_1(t))\,dt \;\middle|\; X_1(0)=x_1\right](1-e^{-\beta T(x_1,x_2,2)})+$$
$$+C(e^{-\beta\tau}-e^{-\beta T(x_1,x_2,2)})+e^{-\beta(T(x_1,x_2,2)+\tau)}(V_0(x_1^\tau,\bar{x}_2,1)-V_k(x_1^\tau,\bar{x}_2,2))\geq 0,$$

which is equivalent to:

$$F_1(x_1,x_2,1):=\frac{\left[\int_0^{T(x_1,x_2,2)} e^{-\beta t}h_2(X_2(t))\,dt \;\middle|\; X_2(0)=x_2\right]}{(1-e^{-\beta T(x_1,x_2,2)})}-$$
$$\frac{\left[\int_0^{\tau} e^{-\beta t}h_1(X_1(t))\,dt \;\middle|\; X_1(0)=x_1\right]}{(1-e^{-\beta\tau})}-C\frac{(1+e^{-\beta T(x_1,x_2,2)})}{(1-e^{-\beta T(x_1,x_2,2)})}+$$
$$\frac{e^{-\beta T(x_1,x_2,2)}}{(1-e^{-\beta T(x_1,x_2,2)})}\left(\frac{e^{-\beta\tau}}{(1-e^{-\beta\tau})}(-C+V_0(x_1^\tau,\bar{x}_2,1)-V_k(x_1^\tau,\bar{x}_2,2))\right)\geq 0 \qquad \text{(F.8)}$$

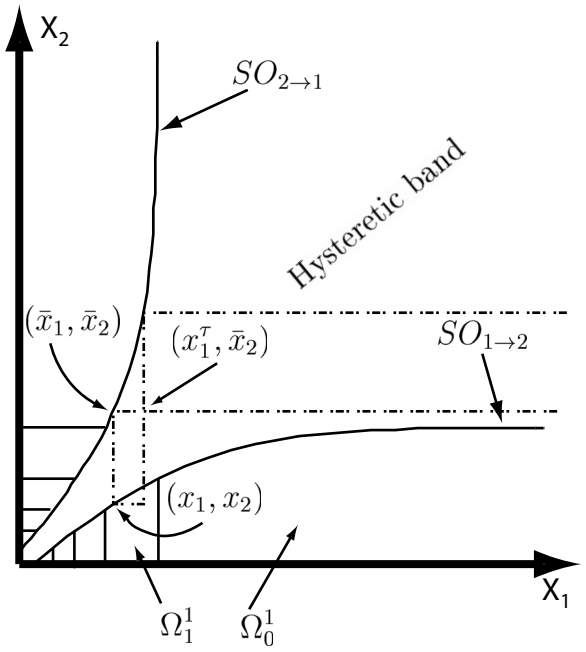

Fig. F.3. Two different realizations.

Lemma F.5. *Given x_1, the function $F_1(x_1,x_2,1)$ (defined in equation (F.8)) is decreasing in x_2.*

Proof:
Remember that $\bar{x}_2$ is the position reached by project 2 at the first switch (i.e. $X_2(T(x_1,x_2,2))=\bar{x}_2$, see Fig.F.3). We note that:

F ⊛ Proof of Proposition 14.4

- The first term is decreasing in x_2. Indeed, making the change of variable

$$y = X_2(t)$$

this first term is rewritten as

$$f_1(x) := \int_{x_2}^{\bar{x}_2} h_2(y) g(y, x_2)\, dy \tag{F.9}$$

where

$$g(y, x_2) = \frac{e^{-\beta X_2^{-1}(y)} \frac{1}{\dot{X}_2}}{\int_{x_2}^{\bar{x}_2} e^{-\beta X_2^{-1}(y)} \frac{1}{\dot{X}_2}\, dy}.$$

By assumption we have that $\dot{X}_2 > 0$ then

$$g(y, x_2) \geq 0 \quad \text{and} \quad \int_{x_2}^{\bar{x}_2} g(y, x_2)\, dy = 1. \tag{F.10}$$

Given $\Psi(x)$ and $\Phi(x, y)$ both of class $\mathcal{C}^1$ in x, it is well known that:

$$\frac{\partial}{\partial x} \int_{\Psi(x)}^{M} \Phi(x, y)\, dy = \int_{\Psi(x)}^{M} \frac{\partial}{\partial x} \Phi(x, y)\, dy - \Psi'(x) \Phi(x, \Psi(x)). \tag{F.11}$$

Using equation (F.11) we have that

$$f_1'(x) = \int_{x_2}^{\bar{x}_2} h_2(y) \frac{\partial}{\partial x_2} g(y, x_2)\, dy - h_2(x_2) g(x_2, x_2).$$

By hypothesis $h_2(x_2)$ is non-increasing then

$$f_1'(x) \leq h_2(x_2) \int_{x_2}^{\bar{x}_2} \frac{\partial}{\partial x_2} g(y, x_2)\, dy - h_2(x_2) g(x_2, x_2). \tag{F.12}$$

Now, from equations (F.11) and (F.10) we have that

$$\int_{x_2}^{\bar{x}_2} \frac{\partial}{\partial x_2} g(y, x_2)\, dy = \frac{\partial}{\partial x_2} \int_{x_2}^{\bar{x}_2} g(y, x_2)\, dy + g(x_2, x_2) = g(x_2, x_2) \tag{F.13}$$

Using equation (F.13) in equation (F.12) we get

$$f_1'(x_2) \leq 0,$$

which prove that the first term of equation (F.8)) is non-increasing in x_2.

- The second term does not depend on x_2.

- The third term is decreasing in x_2 (i.e. its absolute value is increasing).

- The fourth term is decreasing in x_2. Indeed $\frac{e^{-\beta T(x_1, x_2, 2)}}{(1 - e^{-\beta T(x_1, x_2, 2)})}$ is increasing in x_2 (i.e. $T(x_1, x_2, 2)$ is decreasing in x_2) and

$$\Theta = (-C + V_0(x_1^\tau, \bar{x}_2, 1) - V_k(x_1^\tau, \bar{x}_2, 2)) \leq 0.$$

Here, Θ corresponds to the difference between:

i) the reward gained by immediately switching from project 2 to 1 and then by proceeding optimally:

$$-C + V_0(x_1^\tau, \bar{x}_2, 1),$$

ii) the reward gained by staying on project 2 and proceeding optimally with project 2:

$$V_k(x_1^\tau, \bar{x}_2, 2).$$

By construction, $(x_1^\tau, \bar{x}_2)$ lies in the hysteretic band (indeed $(x_1^\tau, \bar{x}_2, 1)$ is on the optimal trajectory starting at $(x_1, x_2, 2)$ see Fig.F.3). In the hysteretic band, the optimal scheduling is to stay on the currently engaged project. Then, we necessarily have:

$$-C + V_0(x_1^\tau, \bar{x}_2, 1) < V_k(x_1^\tau, \bar{x}_2, 2).$$

□ Lemma F.5

Having established that $\digamma_1(x_1, x_2, 1)$ is decreasing in x_2, we can conclude that if $(x_1, x_2, 1) \in \Omega_1^1$ then every $(x_1, x_2', 1)$, with $x_2' < x_2$, also belongs to Ω_1^1. Given x_1, write $\tilde{y}(x_1)$ for the value of x_2 such that $\digamma_1(x_1, \tilde{y}(x_1), 1) = 0$. We then have a function:

$$\tilde{y}(x_1) \ : \ x_1 \mapsto \tilde{y}(x_1).$$

□ Lemma F.4

Lemma F.6 (Iteration n=1, for project j=2). *Assume that* $(x_1, x_2, 2) \in \Omega_1^2$, *then there exists* $\tilde{x}(x_2)$ *such that*

$$\forall z \in]-\infty, \tilde{x}(x_2)], \ (x_1, z, 2) \in \Omega_1^2.$$

Moreover, $\forall (x_1, z', 2)$ *with* $z' > \tilde{x}(x_2)$, *we have* $(x_1, z', 2) \notin \Omega_1^2$.

Proof:
Follow the same arguments as in Lemma F.4.

□ Lemma F.6

Lemma F.7 (Iteration n=n, for project j=1). *Assume that* $(x_1, x_2, 1) \in \Omega_n^1$, *then,* $\exists\ \tilde{y}(x_1)$ *such that,*

$$\forall z \in]-\infty, \tilde{y}(x_1)], \ (x_1, z, 1) \in \Omega_n^1.$$

Moreover, $\forall (x_1, z', 1)$ *with* $z' > \tilde{y}(x_1)$, *we have* $(x_1, z', 1) \notin \Omega_n^1$.

Proof:
By hypothesis $(x_1, x_2, 1) \in \Omega_n^1$, then, using equation (F.2), we have that $\forall \tau \in \mathbb{R}^+$:

$$-C + V_n(x_1, x_2, 2) \geq \left[\int_0^\tau e^{-\beta t} h_1(X_1(t))\, dt \ \middle|\ X_1(0) = x_1 \right] + e^{-\beta\tau} \Big(-C + V_k(x_1^\tau, x_2, 2) \Big). \tag{F.14}$$

Using equation (F.1) and Fig.F.3, this equation can be rewritten in the same form as equation (F.8):

$$\begin{aligned} \digamma_n(x_1, x_2, 1) := & \frac{\left[\int_0^{T(x_1,x_2,2)} e^{-\beta t} h_2(X_2(t))\, dt \ \middle|\ X_2(0) = x_2 \right]}{(1 - e^{-\beta T(x_1,x_2,2)})} - \\ & \frac{\left[\int_0^{\tau} e^{-\beta t} h_1(X_1(t))\, dt \ \middle|\ X_1(0) = x_1 \right]}{(1 - e^{-\beta \tau})} - C \frac{(1 + e^{-\beta T(x_1,x_2,2)})}{(1 - e^{-\beta T(x_1,x_2,2)})} + \\ & \frac{e^{-\beta T(x_1,x_2,2)}}{(1 - e^{-\beta T(x_1,x_2,2)})} \left(\frac{e^{-\beta\tau}}{(1 - e^{-\beta\tau})} (-C + V_{n-1}(x_1^\tau, \bar{x}_2, 1) - V_k(x_1^\tau, \bar{x}_2, 2)) \right) \geq 0 \end{aligned} \tag{F.15}$$

Lemma F.8. *Given x_1, the function $\digamma_n(x_1, x_2, 1)$ is decreasing in x_2.*

Proof:
We can use the same arguments as in Lemma F.5 Indeed,

$$-C + V_{n-1}(x_1^\tau, \bar{x}_2, 1) - V_k(x_1^\tau, \bar{x}_2, 2) \leq 0.$$

This is due to the fact that $(x_1^\tau, \bar{x}_2)$ lies in the hysteretic buffer.

□ Lemma F.8

Having established that $\digamma_n(x_1, x_2, 1)$ is decreasing in x_2, we can conclude that if $(x_1, x_2, 1) \in \Omega_n^1$ then every $(x_1, x_2', 1)$, with $x_2' < x_2$, also belongs to Ω_n^1. Given x_1, write $\tilde{y}(x_1)$ for the value of x_2 such that $\digamma_n(x_1, \tilde{y}(x_1), 1) = 0$. We then have a function:

$$\tilde{y}(x_1) \ : \ x_1 \mapsto \tilde{y}(x_1).$$

□ Lemma F.7

Lemma F.9 (Iteration n=n, for project j=2). *Assume that $(x_1, x_2, 2) \in \Omega_n^2$, then there exists $\tilde{x}(x_2)$ such that*

$$\forall z \in]-\infty, \tilde{x}(x_2)], \ (x_1, z, 2) \in \Omega_n^2.$$

Moreover, $\forall (x_1, z', 2)$ with $z' > \tilde{x}(x_2)$, we have $(x_1, z', 2) \notin \Omega_n^2$.

Proof:
Follow the same arguments as in Lemma F.7.

□ Lemma F.9

We therefore construct by recursion the functions

$$\tilde{y} : \ x_1 \mapsto \tilde{y}(x_1)$$

and

$$\tilde{x} : \ x_2 \mapsto \tilde{x}(x_2)$$

which define the optimal switching curves $\mathcal{SO}_{1\to 2}$ and $\mathcal{SO}_{2\to 1}$ respectively.

□ Proposition 14.4

G

⊛ Proof of Proposition 14.5

From proposition 2, we know that the optimal policy for a two-armed DMABP in the class $\mathcal{Z}$ is described by two switching curves. To end the proof of the claim, it remains to show that these curves are non-decreasing.

Proposition 14.5: The optimal switching curves $\mathcal{SO}_{1\to 2}$ and $\mathcal{SO}_{2\to 1}$, for a two-armed DMABP in $\mathcal{Z}$ are non-decreasing.

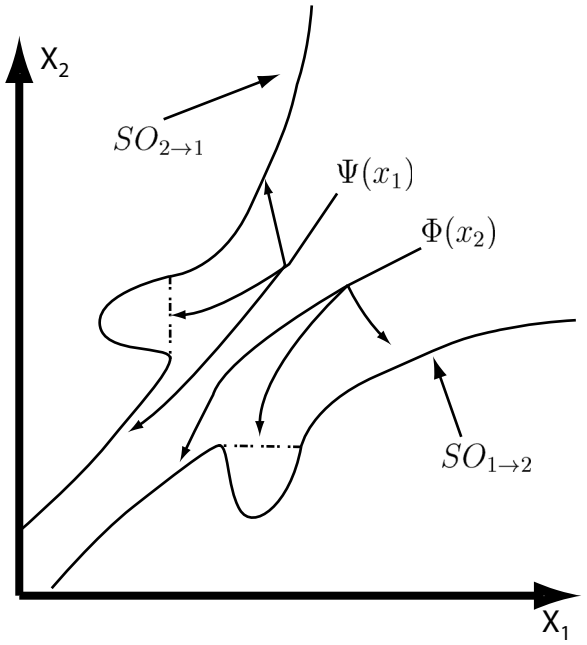

Fig. G.1. Sketch of the functions $\Phi(x_2)$ and $\Psi(x_1)$.

Proof:
Given x_2, let us define $\Phi(x_2)$ to be the smallest value of x_1 such that at $(x_1, x_2, 1)$ it is optimal to immediately engage project 2, namely $\exists\, n \geq 0$ such that:

$$-C + V_n(x_1, x_2, 2) \geq \left[\int_0^\tau e^{-\beta t} h_1(X_1(t))\, dt \;\middle|\; X_1(0) = x_1\right] + e^{-\beta\tau}\Big(-C + V_k(x_1^\tau, x_2, 2)\Big).$$

Then from proposition 14.4 we know that $\mathcal{SO}_{1\to 2}$ is a function hence $\Phi(x_2)$ is non-decreasing in x_2. Similarly, construct the function $\Psi(x_1)$ which is the largest value of x_2

such that at position $(x_1, x_2, 2)$ it is optimal to immediately engage project 1, namely $\exists\, n \geq 0$ such that:

$$-C + V_n(x_1, x_2, 1) \geq \left[\int_0^\tau e^{-\beta t} h_2(X_2(t))\, dt \;\middle|\; X_2(0) = x_2 \right] + e^{-\beta\tau}\Big(-C + V_k(x_1^\tau, x_2, 1)\Big).$$

Again $\Psi(x_1)$ is non-decreasing in x_1 see Fig.G.1.

Lemma G.1. *There exists $u_0 \in \mathbb{R}$ such that the function $\tilde{y}(x_1)$ (i.e. the switching curve $\mathcal{SO}_{1\to 2}$) is non-decreasing on an interval $[u_0, +\infty[$.*

Proof:
By hypothesis the reward $h_1(x_1)$ is decreasing in x_1. Hence, the left hand side of equation (F.4) is also decreasing in x_1. Due to the fact that the righthand side of equation (F.4) does not depend on x_1, we conclude that:

$$\text{if } (x_1, x_2, 1) \in \Omega_0^1 \;\Rightarrow\; \forall\, x_1' > x_1,\; (x_1', x_2, 1) \in \Omega_0^1.$$

Therefore, the function $\tilde{y}(x_1)$ (being the largest value of x_2 such that at $(x_1, x_2, 1)$, the equation (F.4) is satisfied) is non-decreasing on an interval $[u_0, +\infty[$ The position u_0 is the smallest value of x_1 such that equation (F.4) holds (see Fig.G.2).

□ Lemma G.1

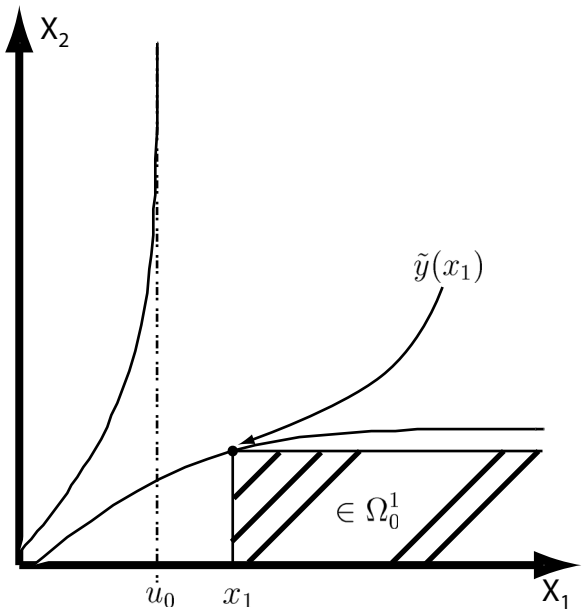

Fig. G.2. The increasing property of the function $\tilde{y}(x_1)$.

Lemma G.2. *There exists $v_0 \in \mathbb{R}$ such that the function $\tilde{x}(x_2)$ (i.e. the switching curve $\mathcal{SO}_{2\to 1}$) is non-decreasing on an interval $[v_0, +\infty[$.*

Proof:
Same proof as for Lemma G.1

□ Lemma G.2

Remarks: Lemma G.1 and Lemma G.2 imply that:

- $\Phi(x_2)$ coincides with $\tilde{y}(x_1)$ on the interval $[u_0, \infty[$.
- $\Psi(x_1)$ coincides with $\tilde{x}(x_2)$ on the interval $[u_0, \infty[$.

Lemma G.3. *The curves $\mathcal{SO}_{2\to1}$ and $\mathcal{SO}_{1\to2}$ are non-decreasing on the interval $[-\infty, u_0[$ and $[-\infty, v_0[$ respectively.*

Proof:
Assume, ad absurdum, that the assertion is false. If so, the optimal policy would locally exhibit the shape sketched in Fig.G.1. To prove Lemma 4.3, we will show that if at a position $B = (p_2, x_2, 1)$ the optimal policy commands to engage the project 1, then $\forall\ A = (p_1, x_2, 1)$ with $p_1 \leq p_2$ the optimal policy would also command to engage the project 1 and hence a contradiction follows (see Fig.G.3).

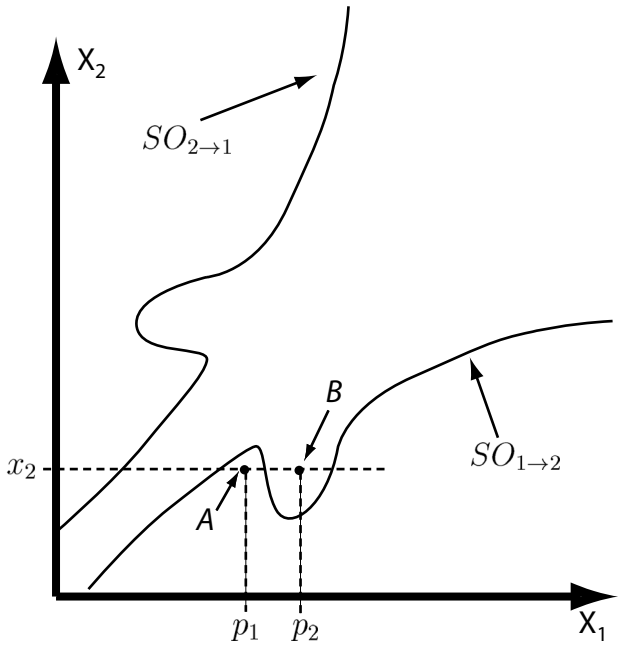

Fig. G.3. The initial conditions A and B.

Let us write

- $\Delta_1(t) = [h_1(X_1(t)) \mid X_1(0) = p_2]$ for the rewards process of project 1 received when engaging it alone with an initial condition $X_1(0) = p_2$ (i.e. starting at B).

- $\tilde{\Delta}_1(t) = [h_1(X_1(t)) \mid X_1(0) = p_1]$ for the rewards process of project 1 received when engaging it alone with an initial condition $X_1(0) = p_1$ (i.e. starting at A).

- $\Delta_2(t) = [h_2(X_2(t)) \mid X_2(0) = x_2]$ for the rewards process of project 2 received when engaging it alone with an initial condition $X_2(0) = x_2$ (i.e. starting at B).

- $\tilde{\Delta}_2(t) = [h_2(X_2(t)) \mid X_2(0) = x_2]$ for the rewards process of project 2 received when engaging it alone with an initial condition $X_2(0) = x_2$ (i.e. starting at A).

G ⊛ Proof of Proposition 14.5

Observe that:

- At both position A and B, the initial conditions of the project 2 is $X_2(0) = x_2$. We then have that $\Delta_2(t) = \tilde{\Delta}_2(t)$.

- The MABP $(X_1, h_1; X_2 h_2)$ is deteriorating. We then have that $\tilde{\Delta}_1(t) \geq \Delta_1(t)$, $\forall\, t \geq 0$.

From these observations is follows that, although the optimal policy $\pi_A(t)$ and $\pi_B(t)$ starting respectively at position A or B may generally differ for $t > 0$, they coincides at $t = 0$. This follows from the fact that starting at B the optimal policy commands to engage $\Delta_1(t)$ at $t = 0$ (this by hypothesis) thus, as $\tilde{\Delta}_1(t)$ is more rewarding $\forall\, t \geq 0$ and $\Delta_2(t) = \tilde{\Delta}_2(t)$ $\forall\, t \geq 0$, it is necessarily more rewarding to engage $\tilde{\Delta}_1(t)$ at $t = 0$ starting as position A.

□ Lemma G.3

Curriculum Vitae

Fabrice Dusonchet
Rte du Pavement 19
1018 Lausanne
Tel : 076 396 05 46
E-Mail : fabrice.dusonchet@epfl.ch

Born: 10 december 1974

EDUCATION

1999:	Master in Mathematics and Computer science, University of Geneva
1999:	Master in Mathematics, University of Geneva
1997:	Academical year at the Mathematics and Computer science department of University of Warwick (England)
1994:	Bachelor of science, Collège de Staël, Geneva

ACADEMIC/TEACHING EXPERIENCE

1999-2003:	Supervise and guide students' master and semester work at the Swiss Federal Institute of Technology - Lausanne
1997-2003:	Teach Mathematics and Computer science to undergraduate students
1999-2003:	Teach Computer science at the adult school of Geneva
2000-2003:	Teach on-going formation in computer science to secondary school teatchers

PUBLICATIONS

- M.-O. Hongler and F. Dusonchet "Optimal stopping and Gittins indices for piecewise deterministic evolution process" Journal of discrete Events Systems 2001 No **11** pages: 235–248
- M.-O. Hongler and F. Dusonchet "Optimal scheduling for piecewise deterministic Multi-Armed Bandit Problem." in Analysis and Modeling of Manufacturing Systems edited by Stanley B. Gershwin, Kluwer November 2002, Hardbound
- M.-O. Hongler and F. Dusonchet "Ordonnancement Dynamique d'une Machine Flexible. Formulation en Processus de Bandits-Manchots" APII-JESA No. **36** (2002)
- F. Dusonchet and M.-O. Hongler. "Continuous Time Restless Bandit and Dynamic Scheduling for Make-to-Stock Production" Accepted for publication by IEEE Trans. on Robo. and Auto. 2003
- F. Dusonchet and M.-O. Hongler "Optimal Policy for Deteriorating Two-Armed Bandit Problems with Switching Costs." Accepted for publication by Automatica 2003

PROCEEDINGS

- F. Dusonchet and M.-O. Hongler. "Continuous Time Restless Bandit and Dynamic Scheduling for Make-to-Stock Production" The Sciences of Complexity University of Bielefeld 2000
- F. Dusonchet and M.-O. Hongler. "Optimal stopping and Gittins indices for piecewise deterministic evolution process" The Sciences of Complexity University of Bielefeld 2000
- M.-O. Hongler and F. Dusonchet "Ordonnancement Dynamique d'une Machine Flexible. Formulation en Processus de Bandits-Manchots" MOSIM'01 (2001) Troyes (France)
- M.-O. Hongler and F. Dusonchet "Dynamic Scheduling of a Multi-Items Production Operating on a Make-to-Stock Basis" ETFA (2001) Antibes (France)
- F. Dusonchet and M.-O. Hongler. "Ordonnancement dynamique optimal en présence de coûts de changement - Un modèle exactement soluble." MOSIM'03 (2003) Toulouse (France)
- F. Dusonchet, M.-O. Hongler and M. Oulevey. "Formule Heuristique pour la Densité Optimale de Palettes à Insérer dans une Chaîne de Production" MOSIM'03 (2003) Toulouse (France)
- F. Dusonchet and M-O. Hongler: "Optimal Policy for Deteriorating Two-Armed Bandit Problems with Switching Costs." CCM (2002) - Universidade da Madeira

References

1. A. Pena-Perez and P. Zipkin. "Dynamic scheduling rules for a multiproduct make-to-stock queue". *Oper. Res.*, 45:919–930, 1997.
2. M. Asawa and D. Teneketzis. "Multi-Armed Bandits with Switching Penalties". *IEEE Trans, on Aut. Cont.*, 41:328–348, 1996.
3. J.S. Banks and R.K. Sundaram. "Switching cost and the Gittins index". *Econometrica*, 62:687–694, 1994.
4. D. Bertsimas and J. Nino-Mora. "Restless Bandits, Linear programming Relaxations and Primal-Dual Index Heuristic". *Oper. Res.*, 48:80–90, 2000.
5. B.I. Grigelionis and A.N. Shiryaev. "On Stefan's problem and optimal stopping rules for Markov processes". *Teor Veroyi Prim.*, 11:611–631, 1966.
6. T. Bielicki and T. R. Kumar. "Optimality of Zero Inventory Policies for Unreliable Manufacturing systems". *Oper. res*, 36:532–541, 1988.
7. C.H. Papadimitriou and J.N. Tsitsiklis. "The complexity of optimal queueing network control". *Math. Oper. Res.*, 24:293–305, 1999.
8. J.F. chipot. "Simuler sans Modéliser". *Industries en Tech..*, 816:62, 2000.
9. M. H. A. Davis. "Piecewise-Deterministic Markov Processes: a General Class of Non-Diffusion Stochastic Model". *J. Roy. Statist. Soc.*, B. 46:353–388, 1984.
10. R. Descartes. *"Discours de la méthode."*. GF Flamarion.
11. F. Dusonchet and M.-O. Hongler. "Ordonnancement Dynamique d'une Machine Flexible. Formulation en Processus de Bandits-Manchots.". *APII-JESA*, 36, pages=117–130,, 2002.
12. F. Dusonchet and M.-O. Hongler. "Continuous Time Restless Bandit and Dynamic Scheduling for Make-to-Stock Production". *Accepted for publication by IEEE Trans. on Robo. and Auto.*, 2003.
13. F. Dusonchet and M.-O. Hongler. "Optimal Policy for Deteriorating Two-Armed Bandit Problems with Switching Costs.". *Accepted for publication by Automatica*, 2003.
14. E. V. Krichagina, S. X. C. Lou, S. P. Sethi and M. I. Taksar. "Production control in a failure-prone manufacturing system: Diffusion approximation and asymptotic optimality". *The Ann. of Appl. Probab.*, 3:421–453, 1993.
15. F. de Véricourt, F. Karaesmen and Y. Dallery. "Dynamic Scheduling in a Make-to stock system: A partial characterization of optimal policies". *Oper. Res.*, 48:811–819, 2000.
16. W. Feller. *"An introduction to probability theory and its applications." Vol II.* J. Wiley, New-York, 1970.
17. S. B. Gershwin. *"Manufacturing Systems Engineering"*. Prentice Hall, 1994.
18. J. C. Gittins. *"Multi-Armed Bandits Allocation Indices"*. J. Wiley, New-York, 1989.
19. J. C. Gittins and D. M. Jones. "A dynamic allocation index for the sequential design of experiments". In J. Gani Ed., editor, *Progress in Statistics*, pages 241–266. North Holland., 1974.
20. P. Glasserman. "Hedging-Point Production Control with Multiple Failure Modes". *Trans. Automat. Cont.*, 40, 1995.
21. H. Kaspi and A. Mandelbaum. " Lévy Bandit: Multi-armed Bandit driven by Lévy processes". *The Ann. of Appl. Probab.*, 5:541–565, 1995.
22. A. Y. Ha. "Optimal dynamic scheduling policy for a make-to-stock production system". *Oper. Res.*, 45:42–53, 1997.
23. Y. Dallery K. H. Youssef and C. Van Delft. "Analyse des Règles de Priorités dans un Système de Pilotage Mixte. Production à la Commande - Production par Anticipation". *APII - JESA*, 36:79–96, 2002.

References

24. I. Karatzas. "Gittins indices in the dynamic allocation problem for diffusion processes". *Ann. Probab.*, 12:173–192, 1984.
25. I. Karatzas and S. E. Shreve. *"Brownian Motion and Stochastic Calculus." Second Edition.* Springer, New-York, 1991.
26. H. Kaspi. Two-armed bandits with switching costs. *Preprint, Technion*, 2002.
27. H. Kaspy and A. Mandelbaum. Multi-armed bandits in discrete and continuous time. *Ann. Appl. Proba.*, 8:1270–1290, 1998.
28. N.V. Krylov. *"Controlled Diffusion Processes"*. Springer Verlag, 1980.
29. L. Shepp and A.N. Shiryaev. "The Russian option : Reduced regret". *Annals Appl. Probab.*, 3:631–640, 1993.
30. S. A. Lippman. "Applying a New Device in the Optimization of Exponential Queuing System". *Oper. Res.*, 23:687–710, 1975.
31. M.-O. Hongler and F. Dusonchet. "Optimal stopping and Gittins' indices for piecewise deterministic evolution process". *J. of Discrete Events Systems*, 11:235–248, 2001.
32. M.-O. Hongler and R. Dalang. "The right time to sell a stock whose price is driven by Markovian noise". *preprint EPFL*, 2003.
33. J. Medhi. *"Stochastic processes"*. J. Wiley, New-York, 1994.
34. J. L. Menaldi and M. Robin. "On the Optimal Reward Function of the Continuation Time Multi-Armed Bandit Problem". *SIAM J. Optim. and Control*, 28:97–112, 1990.
35. M.H Veatch and L. M. Wein. "Scheduling a make-to-stock queue: index policies and hedging points". *Oper. Res.*, 44:634–647., 1996.
36. J. Nino-Mora. "dynamic allocation indices for restless projects and queueing admission control: a polyhedral approach". *Submitted to mathematical programming.*
37. J. Nino-Mora. "Restless Bandit, partial conservation law and indexability". *Adv. Appl. Probab.*, 33:76–98, 2001.
38. M.P. Van Oyen and D. Teneketzis. "Optimal Stochastic Scheduling of Forest Networks with Switching Penalties". *Adv. Appl. Probab.*, 26:474–497, 1994.
39. B. Pascal. *"Pensées."*. Edition Brunschvicg.
40. M.A. Pinsky. *"Lectures on Random Ecolution"*. World Scientific, 1991.
41. M. L. Puterman. *"Markov Decision Processes. Discrete stochastic dynamic programing"*. Wiley-Interscience Publication, 1994.
42. S. M. Ross. *"Stochastic Processes"*. J. Wiley, New-York, 1983.
43. M. Schäl. "Conditions for Optimality in Dynamic Programming and for the Limit of n- Stage Optimal Policies to be Optimal". In Springer-Verlag., editor, *Z. Wahrscheinlichkeitstheorie verw.*, volume 32, pages 179–196. 1975.
44. S. P. Sethi and Q. Zhang. *"Hierarchical Decision Making in Stochastic Manufacturing Systems"*. Birkhäuser, Boston, 1994.
45. A.N. Shiryayev. *"Optimal Stopping Rules"*. Springer, verlag, Applications of mathematic 8., 1978.
46. J. Walrand. *"An introduction to Queueing Network."*. Prentice-Hall International, 1988.
47. R. R. Weber and G. Weiss. Addendum to "On an Index Policy for Restless Bandits". *Adv. Appl. Prob.*, 23:429–430, 1991.
48. L. M. Wein. "Dynamic scheduling of a multi-class make-to-stock queue". *Oper. Res.*, 40:724–735, 1992.
49. P. Whittle. *"Optimization over Time. Dynamic Programming and Stochastic Control." Volume I.* Wiley, New-York, 1982.
50. P. Whittle. "Restless Bandit: Activity in a changing world". In J. Gani Ed., editor, *J. Appl. Probab. A Celebration of Applied Probability*, volume 25A, pages 287–298. 1988.
51. P. Whittle. "Optimal control : Basics and Beyond". In John Wiley and Sons, editors, *Wiley-Interscience series in systems and optimization.* 1996.

Index

Index